nature

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE

TIME TO FIGHT BACK

A new class of antibiotic against multidrug-resistant Gram-negative bacteria **PAGE 189**

METRICS

PROLIFIC AUTHORS

Who publishes a paper every five days?

PAGE 167

ENVIRONMENTAL SCIENCE

PRESERVING WETLANDS

Coastal management might offset effects of sea-level rise

PAGES 183 & 231

PHYSIOLOGY

BONES LAID BARE

Two-way signalling path links formation to resorption

PAGES 180 & 195

NATURE.COM

13 September 2018

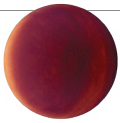
Vol. 561, No. 7722

THIS WEEK

EDITORIALS

PAKISTAN Scientific expertise is essential to shaping the country's future **p.150**

ECONOMICS What the field has learnt, ten years on from Lehman **p.151**



DEATH STARS The stellar pairs that push planets to their death **p.153**

Prior art

The earliest known drawing — crayon on a rock shard — suggests early humans engaged in abstract art.

If a picture tells a thousand words, a cross-hatched design drawn on a fragment of rock some 73,000 years ago could speak volumes. The problem will be understanding what it tells us. The design, reported in *Nature* this week (C. S. Henshilwood *et al. Nature* https://doi.org/10.1038/s41586-018-0514-3; 2018), occurs on a lentil-shaped rock flake, and was found in Blombos Cave, on the southern shore of South Africa, by archaeologist Christopher Henshilwood and his colleagues. The flake bears an abstract design drawn, the authors say, using a crayon made of red ochre.

It is hard to claim that the design is beautiful, dazzling or engrossing. But the artwork is destined to be priceless and famous, because it seems to be the earliest evidence for a drawing in the archaeological record, by some margin. Apart from some cave paintings from Spain dated to around 64,000 years ago — presumably the work of Neanderthals (D. L. Hoffmann *et al. Science* **359**, 912–915; 2018) — the next instance of drawing came around 40,000 years ago with cave paintings found at opposite ends of Eurasia: in the spectacular art decorating the walls of caves in Spain and France, and the more recently discovered cave art in Sulawesi in Indonesia (M. Aubert *et al. Nature* **514**, 223–227; 2014). Despite being located 12,000 kilometres apart, cave paintings such as these contain images that we instantly recognize as figurative art, including a range of animals, and stencils of hands that speak to us, millennia later, as signs of human self-awareness.

A key distinction of this latest piece is that it is a drawing — a design made by applying pigment — rather than an engraving, made by scratching or cutting a design into a surface. Engraving has a longer prehistory than art. The earliest engravings known are on pieces of shell from Trinil, Java, dated to around 540,000 years ago, well before modern humans evolved, and presumably made by *Homo erectus*. Other ancient engravings have been found around the world; all are extremely simple: just lines, sometimes cross-hatched. There is nothing remotely similar to what we would recognize as imagery, and there is insufficient evidence to say whether they might represent something utilitarian, such as tally sticks or calendars. So, were these Palaeolithic hashtags actually designs intended to convey meaning, or mindless graffiti? Some might have been the unintentional result of another action, such as cutting food items, just like the scratches left on a chopping board after slicing a loaf.

A drawing, by contrast, is much harder to dismiss. To be sure, the one from Blombos is as cross-hatched as the engravings, but it could not have been created as the accidental by-product of another process. Although proving intentionality is extremely hard, the authors examine the evidence they have — including detailed study of the ochre residues — with forensic thoroughness. It seems clear that the drawing was a fragment of something bigger, because some of the lines look as if they continued on to pieces now long gone. In addition, the authors attempted to restage history, using pieces of ochre themselves to show that such drawings can be made using crayons carved out of ochre (rather than, say, by brushwork), and that creating the design on such a rock fragment is possible only by deliberate rotation of the design through an angle, much as later artists might rotate their canvas.

That the ancient artist chose to sketch with red ochre is less of a surprise. The mineral, largely consisting of iron oxide, has been used as a pigment since time immemorial. Its earthy red hues clearly meant a lot to the early modern human inhabitants of Blombos Cave and other nearby sites. They used it as an ingredient in paint, and perhaps even as a sunscreen. Between around 100,000 and 73,000 years ago, the people of the region produced artefacts tens of thousands of years in advance of humans anywhere else in the world, including finely worked stone and bone tools and engraved ochre pieces.

> **"It seems clear that the drawing was a fragment of something bigger."**

That the early *Homo sapiens* living there were able to produce such designs suggests they possessed relatively 'modern' cognition and behaviour. What we cannot know is why they made the marks, or what they represent; unlike images of animals or hands, the drawing's abstract nature offers no clues. And that raises a fascinating question about the history of art. Whereas the humans living in South Africa 100,000 years ago were using technology as yet undreamed of elsewhere, they had yet to invent figurative art. So, are the cave paintings of Lascaux and Sulawesi unconnected, independent inventions, or did modern humans create cave art somewhere else along the way, and then take it with them as they moved through the world? What is clear is that they started a trend, one that eventually led to Piet Mondrian, Jackson Pollock, Bridget Riley and the many great artists of today. ■

Noether's legacy

It's time to acknowledge the lasting impact of a brilliant mathematician.

Emmy Noether was a force in mathematics — and knew it. She was fully confident in her capabilities and ideas. Yet a century on, those ideas, and their contribution to science, often go unnoticed. Most physicists are aware of her fundamental theorem, which puts symmetry at the heart of physical law. But how many know anything of her and her life?

A conference in London this week, the Noether Celebration, hopes to change that. It's a welcome move. In a world where young scientists look for inspirational female role models, it is hard to think of a more deserving candidate.

Noether was born in 1882 in Erlangen, Germany. Her parents wanted all their children to get doctorates, so although many universities at the time did not formally accept women, she went. After graduation, sexist regulations prevented Noether from getting jobs in academia. Undaunted, for many years she lectured in Erlangen and, from 1915, at the University of Göttingen — often for free.

At the time, that city was the centre of the mathematical world, largely due to the presence of two of its titans — Felix Klein and David Hilbert. But even when Noether was being paid to teach at Göttingen and making her most important contributions, fate and further discrimination intervened: Hitler took power in 1933 and she was fired for being Jewish. She escaped to the United States and taught at Bryn Mawr College in Pennsylvania, until she died in 1935, at the age of just 53.

Noether devoted her career to algebra and came to see it in a striking new light. "All of us like to rely on figures and formulas," wrote Bartel van der Waerden, her former student, in his obituary of Noether. "She was concerned with concepts only, not with visualization or calculation."

Noether saw maths as what are now called structures. To her, the characteristics of a structure's components — be they numbers, polynomials or something else — mattered less than the networks of relations among an entire set of objects. This enabled her to give proofs that applied to more general structures than the original ones, and which revealed unseen connections.

It was a new and elegant approach that changed the face of algebra. And Noether realized that it could influence other parts of maths. One was topology, a field in which "she published half a sentence and has an everlasting effect", one mathematician wrote. Before Noether, topologists had been counting holes in doughnuts; she brought to bear the full power of her structures to create something called algebraic topology.

The results that Noether published 100 years ago were, for her, a rare foray into physics, in which she was not particularly interested. Albert Einstein had just developed his general theory of relativity, and was struggling to understand how energy fitted into his equations. Hilbert and Klein were working on it, too, and asked Noether for help.

That she did help is an understatement. Noether's expertise in symmetry led her to discover that the symmetries of a physical system are inextricably linked to physical quantities that are conserved, such as energy. These ideas became known as Noether's theorem (E. Noether *Nachr. d. Ges. d. Wiss. zu Göttingen, Math.-phys. Kl.* **1918**, 235–257; 1918).

As well as answering a conundrum in general relativity, this theorem became a guiding principle for the discovery of new physical laws. For example, researchers soon realized that the conservation of net electric charge — which can neither be created nor destroyed — is intimately related to the rotational symmetry of a plane around a point. The impact was profound: those who created the standard model of particle physics, and the researchers who attempt to extend it, think in terms of Noether's symmetries.

"Before Noether, topologists had been counting holes in doughnuts."

Some biographies inaccurately portray Noether as a somewhat helpless genius at the mercy of men's charitable actions. In reality, she was an assertive personality, recognized leader and the first female plenary speaker at the renowned International Congress of Mathematicians.

The status of women in mathematics and science has improved since Noether's time, but bias and discrimination remain. Too few leading female mathematicians receive the recognition they deserve. (Only one woman, Maryam Mirzakhani, has received the Fields Medal, and none has won the Abel Prize — the field's top awards.) Noether is an inspiration: including to UK mathematician Elizabeth Mansfield, who co-organized the London meeting and works on modern extensions of Noether's work.

We don't know how many potential Emmy Noethers have been unfairly denied the chance to show their talents. More people should know — and should celebrate — one who changed the scientific world against the odds. ■

Experts needed

Pakistan needs the courage to listen to independent expertise, including on science.

It took US President Donald Trump 18 months to announce a science adviser. By contrast, Pakistan's new prime minister, Imran Khan, is widely expected to appoint a science minister in the early stages of his administration. It will be a welcome move: scientific expertise will be essential in shaping the country's future. But whoever it is will have a fight on their hands to ensure that inclusive and evidence-based advice prevails.

Khan has come into office promising anything but business as usual. In an echo of America's New Deal and the first post-Second World War European governments, he has pledged a welfare state, affordable health care, school reform, improvements to agriculture (the backbone of the economy), and an axe to public-sector corruption.

To do all this, his government will need medical researchers and health-care specialists to advise on plans for a national health service; primary- and secondary-school educationalists to work out how best to get every child into a good school; and science and innovation policy experts to guide academic researchers on the path to quality improvements and better community engagement.

Khan has repeatedly said that he will appoint the best people to top jobs. On that score, however, his ministerial team of mostly political appointees has had mixed reviews. What the new ministers lack in particular is a wide-ranging and credible network of experts to call on. There is a shortage of people — and especially women — from high-level academic and other professional backgrounds plugged into policymaking. Khan knows this and has appealed for help. So far, the calls have been heeded by economists in particular, as evidenced by the prime minister's 18-member Economic Advisory Council — albeit an all-male one.

But one obstacle to gathering expertise is the state's persistent failure to confront rising intolerance, particularly against minorities. In a backwards move, Khan last week bent to the will of the TLP, a far-right political party that attracted more than 2 million votes in July's elections, and he removed Princeton University economist Atif Mian from the Economic Advisory Council. Mian's nomination was challenged by the TLP solely because he belongs to the Ahmadiyya, a much-persecuted minority Muslim community. Mian is highly regarded and his dismissal has been roundly condemned; the other two international members of the council resigned in protest.

The move signals an unwillingness by Khan to appoint advisers who can speak truth to power — and that could have a wider, chilling effect. It will make other independent experts think twice about joining Khan's cause or advocating vital but unpopular reforms.

The new government has ratcheted up expectations with a list of tasks on which lives depend. One of Pakistan's most urgent challenges, for example, is to improve the availability and quality of water. Agriculture uses 90% of supplies, but a population of 200 million — and rising — means that the country is officially classified as 'water-scarce'. Climate change is projected to reduce water availability further, and poor water quality is a major source of disease. In response, the government wants to build more dams, but the expert consensus is that the cost of dams outweighs the benefits, and so other solutions must be sought.

Will that type of expertise be listened to, and will Pakistan forge the right path? Right now, citizens at home — and many in the international community — are willing Imran Khan's nation-building project to succeed. As a star cricketer, Khan once described his style of play as that of a cornered tiger. He needs to muster this legendary courage and call out bigotry; otherwise, whatever goodwill exists towards his government will evaporate very fast indeed. ■

WORLD VIEW *A personal take on events*

LIAM ACTON

Post-crash economics: have we learnt nothing?

A decade on from a worldwide financial meltdown, economics teaching is still stuck in the past, warns **Maeve Cohen**.

Ten years ago, the behemoth investment bank Lehman Brothers collapsed. Its fall was the most high-profile symptom of the failings of economics practitioners, which collectively led to a global financial crisis. In the aftermath, countless articles, books and films were created outlining the hubris of economists and their inability to see the failings of their discipline. This led to much soul searching, defensiveness and disbelief.

A decade later, core questions still stand: was such a high level of confidence ever warranted? Why weren't economists able to predict or stave off the crash? How had they steered the world from prosperity into austerity?

Growing up in an old mining village in the north of England, I saw the consequences of economic events on ordinary lives. I left school at 16 and, when the crash came, I was funding a life of travel by working in cafes and bars. The collapse put an end to this lifestyle. I could no longer earn enough. After my father's business went under, I decided I needed to understand more about how our economic system works.

I went to university and was stunned by the irrelevance of what I was being taught. It was as if the crash hadn't happened. If this was how we were training our economists, it was no wonder our economic system wasn't working! My fellow students and I started a campaign for curriculum reform. In 2014, we released a report analysing the failings of the economics curriculum at our university, with a foreward by the chief economist of the Bank of England (see go.nature.com/2cqrifz). Within hours of its release, it had been downloaded more than 10,000 times. Dissatisfied students from around the world came together, eventually creating an organization called Rethinking Economics, which I now head.

> I WAS **STUNNED** BY THE **IRRELEVANCE** OF WHAT I WAS BEING **TAUGHT.**

Our book, *The Econocracy* (2016), surveyed 174 economics modules at 7 leading UK universities and found that fewer than 10% covered anything other than mainstream economics. Students in Denmark, the Netherlands and Norway have done similar reviews, with similar results. Undergraduate economists all over the world learn theories from textbooks that have barely changed since the 1950s. Those theories are based on individual agents, competing in markets to maximize narrowly defined 'economic utility' (for people) or profit (for firms). The principles are taught with the same certainty as Newtonian physics, and are as devoid of value judgements.

This is absurd. Clearly, there are values; mainstream economics values efficiency, markets and growth, and puts individuals over collectives. Yet, undergraduates are not taught to recognize, let alone question, these values — and the consequences are serious. Economists might disagree on specific predictions, but there are few disagreements about the methodology and the values it assumes.

I believe academic economists are increasingly aware of the shortcomings of their discipline. Although growth and markets remain central themes, students who enter graduate school are trained to re-examine and critique assumptions. Most realize that although their methods might be precise, consistent and derived from empirical data, social behaviour is messy and contextual. Few would argue that economics is now, or ever will be, on the same level as the natural sciences.

This humility does not necessarily extend to those who leave university after their degree, as the vast majority of economics graduates do. These are the people who go on to work in big business, governments and central banks, who shape policy and create our 'economic common sense'. So it was in 2008; so it still is today.

To build a better society, we need undergraduate education that is pluralist, critical and relevant to society. Students should learn mainstream economics, but it should be juxtaposed with the myriad other schools of economic thought, such as ecological economics (which embeds the economy in the environment, rather than thinking of it as an externality), feminist economics (in which gender relations, and unpaid domestic labour, are integral to how the economy functions) or post-Keynesian economics (which proposes, among other things, that the size of an economy is determined by what people are able to buy, rather than by what they produce). Teaching multiple schools of thought will reveal the assumptions, benefits and weaknesses of various theories. Students should also learn economic history, and have enough knowledge of politics and philosophy to contextualize policy outcomes, or at least to understand that the Great Depression of the 1930s influenced Keynes's thinking. They need to gain experience in handling real-world data and recognizing the strengths and shortcomings of data sets and sources.

What gladdens me is that the conversation about economics education has shifted. Many in the academic mainstream now acknowledge that there is a problem in undergraduate teaching. Unfortunately, the efforts and incentives for fixing it are still much too weak.

Academics are rewarded for publishing in top journals, not for revising core curricula. And to get hired by a leading university in the first place, their research must be published in a top journal, which focus mainly on mainstream economics. So even economics departments that want to create a more diverse curriculum end up ill-equipped to do so.

If we are to move beyond the dead hand of antiquated economics, we must create rewards, grants and accolades for training undergraduates better. It is time we taught economics students to practice humility, scepticism and caution from the very beginning. ■

Maeve Cohen *is the director of Rethinking Economics in Manchester, UK, a non-profit group devoted to improving economics education.*
e-mail: maeve.cohen@rethinkeconomics.org

SEVEN DAYS *The news in brief*

SPACE

Exoplanet leads

Astronomers are studying the first batch of possible planets spotted by NASA's Transiting Exoplanet Survey Satellite (TESS), which hunts for worlds around nearby bright stars. The 'TESS Objects of Interest', released to other scientists for follow-up observations on 5 September, include 73 objects. Researchers think that 5–20% of them will probably turn out to be false alarms. Some of the objects are planets already known from ground-based searches. But others, if confirmed, will join the ranks of newly discovered exoplanets. TESS launched in April and began its two years of science operations in July. It is currently searching for planets in the southern sky; the latest results come from the first slice of that survey. See go.nature.com/2owq163 for more.

ALASTAIR POLLOCK PHOTOGRAPHY/GETTY

Australian fur-seal populations in decline

Numbers of Australian fur-seal (*Arctocephalus pusillus doriferus*) pups have declined for the first time in more than three decades, according to a study published on 5 September (R. R. McIntosh *et al. PLoS ONE* **13**, e0200253; 2018). Researchers compared the latest count, collected in 2013–14, with an overall trend in the population since monitoring began in 1986. Pup numbers roughly doubled between 1986 and 2002, to reach about 20,000 individuals. The population increased slightly between 2002 and 2007. But only 16,516 pups were recorded during a population-wide survey of colonies along the coasts of southeast Australia in 2013–14. This represents a decline of around 4% per year since 2007. There was some good news. Three new colonies were identified in 2013. The researchers suggest that overall pup numbers could be declining because some colonies are overcrowded and food availability was unusually low in 2013–14. But they also raised the concern that this could be the start of a sustained decline in Australian fur-seal numbers.

RESEARCH

Gene-therapy hope

A therapy that edits genes directly in the human body might be safe, suggest early findings from the first trial to test the approach. Researchers from Sangamo Therapeutics in Richmond, California, designed enzymes to correct an error in the genome of people with a rare genetic disease called Hunter syndrome, who cannot break down some complex sugars. The scientists used a virus to deliver these gene-editing enzymes into the cells of patients. Sangamo presented the results on 5 September at the annual meeting of the Society for the Study of Inborn Errors of Metabolism in Athens, Greece. The data on the efficacy of the approach come from just 4 people enrolled in the study for 16 weeks. Although there are some signs that the treatment reduced levels of a biochemical marker used to assess the severity of Hunter syndrome, it is too soon to know whether the therapy worked.

Election security

Paper ballots that can be tallied by hand are the most secure way to conduct an election, according to a report released on 6 September by the US National Academies of Sciences, Engineering, and Medicine. The analysis calls for all US elections to be conducted using such ballots by the 2020 presidential election. It comes after US intelligence agencies concluded that the Russian government backed attempts to infiltrate the United States' election infrastructure during the 2016 presidential election. The report's recommendations were developed by a committee whose members had experience ranging from computer science to officiating at elections.

PEOPLE

US science adviser

Meteorologist Kelvin Droegemeier is one step closer to becoming director of the White House Office of Science and Technology Policy (OSTP) and science adviser to President Donald Trump. The Senate Committee on Commerce, Science, and Transportation voted unanimously on 5 September to approve Droegemeier. The scientist, whom Trump nominated on 31 July, now faces one last hurdle: a vote by the full Senate, which is expected to take place later this month. Droegemeier, an expert on extreme-weather events, recently stepped down as vice-president for research at the University of Oklahoma in Norman. He was a member of the National Science Board, which oversees the National Science Foundation, during the administrations of former presidents George W. Bush and Barack Obama. If confirmed, Droegemeier would be the first non-physicist to serve as White House science adviser since Congress established the OSTP in 1976.

ENVIRONMENT

Australian fires

Large parts of southern Australia are facing a greater than usual risk of bush fires owing to drought conditions and above-average temperatures. The latest seasonal forecast predicts high risk levels for more than one-quarter of the eastern states of New South Wales (NSW) and Victoria, particularly coastline regions, from August until March. The entire state of NSW was declared to be in drought in July (**pictured**). This year's rainfall figures for the state are among the lowest 10% recorded for more than a century. The states of Victoria, Queensland and South Australia have also all experienced lower-than-average rainfall. "We're seeing moisture levels in the vegetation and soils that we haven't seen this bad for 20 years or more," says Richard Thornton, chief executive of the Bushfire and Natural Hazards CRC in East Melbourne, Australia, which publishes the bush-fire forecasts. Southern Australia has also experienced above-average daytime maximum temperatures for 22 consecutive years.

DAVID GRAY/REUTERS

FUNDING

Japan budget

Japan's science ministry has requested a budget increase of 21% for 2019, totalling ¥1.17 trillion (US$10.58 billion). Under the proposal, the country's next-generation supercomputer, Post-K, which is under development at RIKEN in Kobe, would receive ¥20.6 billion, more than triple its 2018 budget. As Japan's first exascale supercomputer, Post-K will be able to carry out a billion billion calculations per second. Spacecraft development and space science would receive ¥153 billion, a boost of more than 30%. The budget also earmarks ¥43.1 billion for the international Thirty Meter Telescope (TMT) — Japan is a member of the TMT consortium — which is awaiting approval for construction in Hawaii, and a feasibility study for the proposed next-generation neutrino detector, Hyper-Kamiokande. The budget is likely to be pared down before final government approval, expected in December; Japan's cabinet is aiming to cut discretionary spending by 10%.

POLICY

Indian council

India has a new science, technology and innovation council that will advise its prime minister. Geneticist Krishnaswamy VijayRaghavan, who is also principal scientific adviser to the government, will lead the council of nine members. The heads of 11 science departments will be 'special invitees' to the council. Unlike the country's previous scientific advisory groups, the panel will go beyond having an advisory role to work with the government to develop and implement policies, VijayRaghavan told *Nature*. Its priorities include encouraging industry — historically averse to investing in scientific research — to collaborate with research institutions. It will also aim to make science and technology more inclusive by implementing programmes in both English and local languages.

EVENTS

CRISPR patents

A fierce and unprecedented patent battle between two educational institutions might be nearing a close, after a US appeals court issued a decisive ruling on the rights to CRISPR–Cas9 gene editing. On 10 September, the US Court of Appeals for the Federal Circuit awarded the pivotal intellectual property to the Broad Institute in Cambridge, Massachusetts, upholding a previous decision by the US Patent and Trademark Office. The decision spells defeat for a team of inventors at the University of California, Berkeley, led by molecular biologist Jennifer Doudna. See go.nature.com/2mkprnc for more.

TREND WATCH

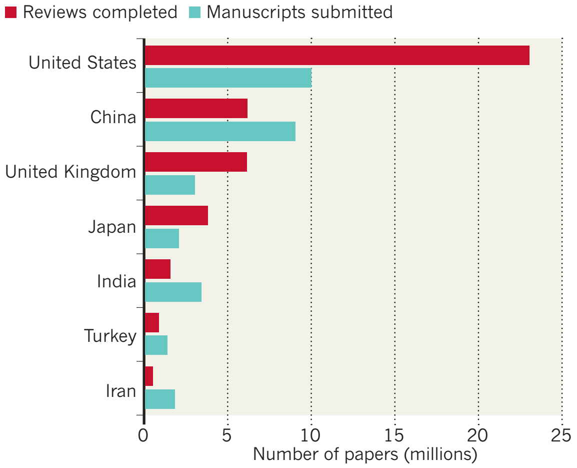
Scientists in developed countries provide nearly three times as many peer reviews per paper submitted as researchers in emerging nations, according to the largest ever survey of the practice.

The Global State of Peer Review report, published on 7 September, was undertaken by Publons, a website that helps academics to track their reviews and other contributions to journals. The authors used data from their survey of 11,000 researchers worldwide, conducted from May to July this year, as well as data from the Publons, Web of Science Core Collection and Scholar One Manuscripts databases.

Although contributions to peer review from emerging economies are lower than those of developed countries, they are rising rapidly, says Andrew Preston, managing director of Publons, in London. "Peer reviews lag publication, so it will take a few years for emerging regions to catch up," he says. They found that, in 2017, editors had to invite, on average, 2.4 scientists to find a reviewer — up from 1.9 invitations in 2013. The report also notes that finding willing peer reviewers is becoming harder, even as the overall volume of publications rises globally. See go.nature.com/288knde for more.

UNEVEN CONTRIBUTIONS

Researchers in the United States and the United Kingdom tend to review more papers than they submit, whereas those in China and India review fewer.



SOURCE: R. R. MCINTOSH *ET AL. PLOS ONE* **13**, E0200253 (2018).

NEWS IN FOCUS

POLICY South African science gets biggest shake-up in 20 years **p.158**

TECHNOLOGY Google targets scientists and data geeks with new search tool **p.161**

AWARDS Astrophysicist to donate US$3 million to promoting diversity **p.161**

ENERGY Data centres try to keep electricity use from skyrocketing **p.163**

FRED GREAVES/REUTERS

Firefighters in California have faced a historic fire season in 2018, and climate models predict that wildfires will become more common and intense.

EPIDEMIOLOGY

Scientists scramble to study wildfires' health effects

Blazes have created natural experiments in nearby towns and a monkey-breeding colony.

BY SARA REARDON

Record-setting wildfires have burnt through northern California over the past month, blanketing huge swathes of the western United States in a smoky haze and destroying an area larger than London. Now scientists are hoping that the fiery summer will help them determine whether exposure to wildfire smoke damages health over the long term.

Finding answers is becoming more urgent because the behaviour of wildfires — in the United States and elsewhere — is expected to shift in the coming decades. Climate models predict that many more people worldwide will be exposed to toxic smoke as these blazes become more common and intense. US wildfires already produce about one-third of the country's particulate-matter pollution, airborne particles that are small enough to enter and damage human lung tissue[1].

"When we think about climate-change policy and cost–benefits, if we don't include human-health impacts we're not getting an accurate assessment," says Michelle Bell, an environmental-health researcher at Yale University in New Haven, Connecticut. "The line between natural and anthropogenic air pollution has blurred in terms of wildfires."

One of the reasons researchers know so little about wildfires' effects on health is that epidemiological studies of air pollution typically do not distinguish between the sources of pollutants that people breathe in. The sensors used in such research measure only the size of particles in the air, making it hard to link ▶

▶ a specific health impact to diesel exhaust, wood smoke or any other source of pollution. That is especially true of wildfires, which consume trees, buildings, synthetic materials and anything else in their path.

"It is a crazy mix of chemicals," says Christopher Migliaccio, an immunologist at the University of Montana in Missoula. "You have no idea how that mixture interacts within the human body to know what the real culprit is."

But Migliaccio is attempting to sort out just that. In 2017, a wildfire 13 kilometres from Seeley Lake, Montana, exposed the town of 1,600 people to nearly 30 times the level of particulate matter considered safe by the US Environmental Protection Agency (EPA). The fire burned for 70 days. After it subsided, Migliaccio and his colleagues collected blood and measured respiratory function in 100 volunteers, whom they plan to track for years.

Early data from the study suggest that the fire harmed the people's lungs and immune systems, Migliaccio says, and those effects persist even a year later. The researchers are now collecting data from residents of two other towns in western Montana. Historical trends suggest that at least one of the three towns will be exposed to a wildfire each year, which will allow the team to compare participants' physiology and health before and after a fire.

Another group of scientists — this one at the University of California in Davis (UC-Davis) — is tracking the physical and mental health of people who were exposed to a series of 250 fires last year in California's Napa and Sonoma valleys that caused US$85 billion in damage. The scientists are going door-to-door to recruit volunteers, and more than 2,000 households have responded, says UC-Davis environmental epidemiologist Irva Hertz-Picciotto. The team is also collecting placentas and cord blood from women who were pregnant during the fires, to determine the chemicals to which their babies might have been exposed.

Wildfire smoke could harm young monkeys.

CALIFORNIA NATIONAL PRIMATE RESEARCH CENTER

Researchers are also beginning to untangle how the composition of material burnt during a wildfire affects the body. Smoke from burning pine needles damaged the DNA of mice in a recent EPA study more than did smoke from burning plastic; burning eucalyptus was the most toxic to immune cells found in the animals' lungs[2]. The UC-Davis researchers are collecting ash given off by the 2017 fires in northern California to analyse its chemical composition and to look for links to specific health effects in their volunteers.

Across town, at the California National Primate Research Center, the results of a natural experiment could provide crucial data on fires' effects on children. In August, smoke from the Mendocino Complex fire — a blaze that has burnt some 186,000 hectares in California — drifted 320 kilometres south to the research centre, which is home to an outdoor colony of primates bred for research. About 2,000 animals, including roughly 500 infant rhesus macaques, were exposed to potentially toxic smoke levels over 10 days.

Lisa Miller, a respiratory immunologist at UC-Davis, and her colleagues have collected blood samples and other data from the infant monkeys. They're looking for links between smoke exposure and the long-term respiratory damage that is often seen in children exposed to air pollution. "We knew for decades that early exposure results in long-term changes" to kids' lung function, Miller says. "What we don't understand, particularly for paediatric populations, is what exactly changed."

The group's previous data suggest that exposure to wildfire smoke can harm young monkeys. In 2008, animals at the centre were exposed to smoke from a wildfire 320 kilometres away. Animals that were infants at the time ended up with smaller, stiffer lungs and weaker immune systems than those of monkeys born the next year[3].

Once the researchers determine how exposure to smoke harms the lungs and immune systems in people and monkeys, it could become easier to prevent or treat the damage. There is no time to waste, says Hertz-Picciotto. "Now that this is becoming more common, it should be possible to prepare and reduce the amount of suffering." ■

1. Rappold, A. G. *et al. Environ. Sci. Technol.* **51**, 6674–6682 (2017).
2. Kim, Y. H. *et al. Environ. Health Perspect.* **126**, 017011 (2018).
3. Black, C. *et al. Am. J. Respir. Cell. Mol. Biol.* **56**, 657–666 (2017).

RESEARCH POLICY

South Africa pushes science to improve daily life

Sweeping policy changes aim to refocus research efforts to tackle issues such as poverty.

BY SARAH WILD

South Africa's science system is set for its biggest shake-up in 20 years, amid proposed legislation changes that aim to make research efforts better serve citizens and address problems such as poverty and unemployment.

Policymakers in the government's department of science and technology are updating the 1996 legislation document that governs the country's science, technology and innovation activities and agencies. A final draft, seen by *Nature*, shifts the focus of South Africa's science sector towards business-led innovation that tackles societal problems and expands the economy. It also reaffirms a key government goal of boosting total research and development (R&D) spending from 0.8% to 1.5% of gross domestic product (GDP) in the next decade. On 5 September, the document was approved by the government's cabinet. It will now be referred to Parliament, where it will be open for public comment before being signed into law.

"If we don't make an impact on the lives of South Africans, then we don't deserve to exist," says the country's science minister, Mmamoloko Kubayi-Ngubane, of her department. Kubayi-Ngubane has overseen the drafting of the latest legislation and was appointed this year by President Cyril Ramaphosa, who took over in February and is widely seen as more pro-science

MUJAHID SAFODIEN/AFP/GETTY

Science minister Mmamoloko Kubayi-Ngubane is overseeing revisions to South Africa's research policy.

than his predecessor, Jacob Zuma. During Zuma's nine-year tenure, the nation's currency weakened in value, and science budgets struggled to keep up with inflation.

Still, South Africa is home to one of the continent's strongest science systems. In part, that is because the apartheid government, which was in power until 1994, focused on military R&D in an effort to uphold white-minority rule and circumvent international economic sanctions at the time.

In 2015–16, government and private sources together spent about 32 billion rand (US$2 billion) on research (see 'Science in South Africa'). The government's science budget accounted for about 14.4 billion rand of that total, supporting 26 public universities and almost 52,000 researchers — figures comparable to those for a highly developed country such as Norway, although South Africa's population is about 10 times larger.

In recent years, South Africa has also positioned itself as an international player in astronomy, taking advantage of regions with clear skies. Its role in the Square Kilometre Array, a project to build the world's largest radio telescope, partly in South Africa, has bolstered its publications and citation impact in astronomy and astrophysics.

BUSINESS BOOST

But R&D spending is still nowhere near the 1.5% level, says Kubayi-Ngubane, which is why the new policy aims to boost business spending on research. "Government can't do it alone. We need the private sector to come in and spend on R&D," she says. Countries in the Organisation for Economic Co-operation and Development spend, on average, about 2.3% of their much larger GDPs on R&D. In these nations, business spending on R&D outstrips government spending, whereas in South Africa, the opposite is true.

South Africa already has initiatives for attracting business and promoting R&D, but companies have remained reticent. For example, since 2006, the government has offered firms a tax incentive to attract business R&D, but uptake has been slow. To improve matters, the draft legislation specifically instructs the department of science and technology to organize a biannual conference to bring together leaders in business, government, higher education and civil society to identify problems in the system and chart a way forward.

The revised legislation also aims to improve other elements of the existing policy. A 2016 review of the original legislation noted that it had been successful in setting up institutions and transforming the demographics of the country's researchers, which previously comprised mainly white scientists, and that publication output had tripled since 1996. But it found that science and innovation had had a limited impact on the country's high unemployment rate, which stands at about 27%. Patent output was also low, and research efforts had yielded few marketable technologies.

The proposed legislation now puts explicit goals on producing more university graduates, publications and patents. And on top of large projects, it lays out areas of innovation that it plans to exploit. One is big data. As part of its astronomy push, South Africa has invested heavily in big-data infrastructure, an investment that could, for example, create business opportunities and promote electronic governance platforms to speed up the delivery of services to citizens.

The document also highlights the green economy, which ranges from recycling to clean-energy research, as an avenue for economic development, and as a way of mitigating and adapting to climate change. South Africa has been experiencing a crippling drought, which this year saw Cape Town's water reserves almost run dry. "Today we deal with issues of drought," says Kubayi-Ngubane. "So what are the new technologies? Are scientists able to share with us new plans?"

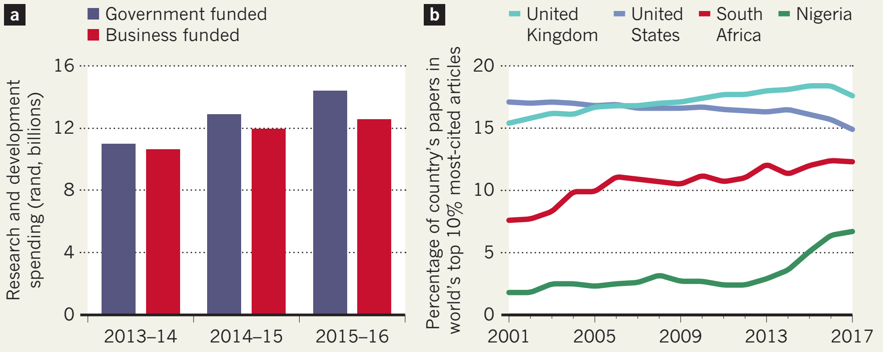
ACADEMIC VERSUS APPLIED

John Mugabe, a science-policy specialist at the University of Pretoria in South Africa, says that the 1996 legislation had similar goals for economic growth, but the government has struggled to translate them into action. He suggests that new legislation should differentiate between science policy and technology policy, and thinks that there should be more emphasis on the latter. "Economic development will only come through enhanced industrial productivity in manufacturing, agriculture and mining," he says.

The perennial question in South Africa, which has many immediate demands on the national purse, is how to fund its science ambitions. But science and innovation is fundamental to strengthening South Africa's economy, says Kubayi-Ngubane. "Where do we want South Africa to be in 20 years? Because science, technology and innovation will influence where we will be." ■

SCIENCE IN SOUTH AFRICA

The South African government spent about 14.4 billion rand (US$940 million) on research and development in 2015–16 (**a**). Business spending lags behind, at 12.6 billion rand. Around 12% of South African-authored research papers are now in the world's top 10% of most-cited articles (**b**).



SOURCES: SOUTH AFRICAN NATIONAL SURVEY OF RESEARCH AND EXPERIMENTAL DEVELOPMENT: STATISTICAL REPORT 2015/16 (LEFT); SCIVAL (RIGHT)

HUMAN SPACEFLIGHT

India's surprise plan to send people to space by 2022

Expensive mission could divert scarce resources from more useful space programmes.

BY SANJAY KUMAR

India has set an ambitious goal to send people to space by 2022. The mission and its short timeline — announced by Prime Minister Narendra Modi— surprised many, including the head of the country's space agency.

"It came as a surprise to us," said Kailasavadivoo Sivan, chair of the Indian Space Research Organisation (ISRO). Although Indian space enthusiasts have discussed sending people into orbit for more than a decade, the idea has not gained major political support until now.

Following Modi's announcement on 15 August, Sivan said that his agency can achieve the task with the help of other nations. ISRO is now working on a plan to send three astronauts to space for between five and seven days, he said. It will prepare for that mission with two uncrewed missions that will attempt to reach a low Earth orbit of about 300–400 kilometres. The government says the entire programme will cost about 100 billion rupees (US$1.4 billion).

"It is possible to do a very basic programme in a short period of time," says John Logsdon, a space-policy specialist at George Washington University in Washington DC. If successful, India would become the fourth country to launch its own manned spacecraft, after the United States, Russia and China.

> *"Human spaceflight with adequate risk cannot be inexpensive."*

But some observers say that India has a lot of work ahead of it and might struggle to meet the deadline. Others have criticized the government for announcing a space programme without consulting ISRO. "It would be ludicrous not to look at Modi's pronouncement as an upcoming 2019 election speech rather than as a road map for India's space programme," says Gauhar Raza, a retired researcher and former chief scientist of the Council of Scientific and Industrial Research in New Delhi.

ISRO already has several of the major components needed for a crewed spacecraft, says Sivan. These include a module to carry astronauts and a crew escape system, which would be used in the event of a mission being aborted mid-launch. ISRO has developed and successfully tested a reusable space shuttle, which could transport and return the astronauts. The agency's Geosynchronous Satellite Launch Vehicle (GSLV) Mark III rocket will probably be used to launch the craft.

CAN IT BE DONE?

The agency will have to upgrade some of its technologies, and test them all for crewed missions, says Ajey Lele, a senior fellow with the Institute for Defence Studies and Analyses, a government-funded think tank in New Delhi. For instance, the GSLV Mark III rocket will need to be able to launch crewed spacecraft, which are much heavier than the satellites the rocket is designed to carry, he says.

Although the agency has developed technologies to support humans in space, such as spacesuits and an environment-control and life-support system, it has no experience in keeping astronauts healthy while in space, or returning them safely, says Lele.

Sivan says ISRO will need the help of other institutions — such as the Institute of Aerospace Medicine in Bangalore — and more-experienced countries to develop some of these capabilities. "We will need some outside cooperation and that may reduce our time and save us money," he says. To train its astronauts, India will probably have to use facilities in the United States, Russia or Europe, says Gopalan Madhavan Nair, who was chair of ISRO from 2003 to 2009.

Logsdon says that India could also buy technologies from commercial operators while it develops its own equivalents. But he warns that it will be costly. "Human spaceflight with adequate risk cannot be inexpensive," he says.

Some science-policy analysts have been highly critical of the human space programme. Prabir Purkayastha, vice-president of the non-profit All India People's Science Network, which promotes the use of science to serve the public good, says that most of ISRO's projects cater for India's development needs, for example by launching satellites for communication and weather forecasting. Sending people to space "is essentially a vanity project that could divert scarce resources and derail ISRO's existing programmes", he says. ■

The Indian space programme has launched many satellites, but now it aims to send people into orbit.

ARUN SANKAR/AFP/GETTY

DAVID HARTLEY/SHUTTERSTOCK

Jocelyn Bell Burnell discovered pulsars as a PhD student at the University of Cambridge, UK.

BREAKTHROUGH PRIZE

Pulsar discoverer wins $3-million prize

Jocelyn Bell Burnell to use cash to promote diversity in science.

BY ZEEYA MERALI

Fifty years after discovering pulsars — compact rotating stars that emit beams of radiation — astrophysicist Jocelyn Bell Burnell has been awarded one of the most lucrative prizes in science: a US$3-million Breakthrough prize. Thought by many to have been snubbed for a Nobel prize for the discovery, Bell Burnell, 75, has been recognized by the Breakthrough committee with a special award in fundamental physics both for her scientific achievements and for her “inspiring leadership” over the past five decades.

“I cannot think of a more deserving scientist to win this prize,” says Chiara Mingarelli, an astrophysicist at the Flatiron Institute in New York City. “In addition to being both a pioneer and a giant in the field, Bell Burnell is the highest calibre role model — a champion for women in science.”

Bell Burnell, now at the University of Oxford, UK, and the University of Dundee, UK, was “totally surprised” to learn about the prize, which was announced on 6 September. “For once in my life, I was speechless,” she says. She is already in discussions with the national physics institutes in the United Kingdom and Ireland about using the prize money to create PhD studentships for people from under-represented groups in science. “Diversity is very important,” says Bell Burnell. “This also recognizes that I did my most important work as a student.”

The Breakthrough prizes were launched in 2012 and are funded by entrepreneurs including Google co-founder Sergey Brin. Awarded in fundamental physics, life sciences and mathematics, they are usually handed out in December, based on selections made after an open nomination process. But the selection committee can decide to make special awards. Previous special awards have been given to Stephen Hawking and to the Laser Interferometer Gravitational-Wave Observatory collaboration for the discovery of gravitational waves.

Pulsars are dense stars, consisting mostly of neutrons, that rotate at a precise rate, emitting radiation as they spin. In 1967, Bell Burnell, then a PhD student at the University of Cambridge, UK, under astronomer Antony Hewish, was analysing hundreds of metres of chart paper containing data collected by a radio telescope in Cambridge when she noticed some mysterious recurring smudges. She was able to characterize these as signs of radio pulses emanating from a spinning star: the pulsar. “The discovery is a testament to her curiosity, her determination and her creativity,” says Mingarelli.

In 1974, Hewish shared the Nobel Prize in Physics with fellow radio astronomer Martin Ryle, for pioneering research in astrophysics. Hewish was cited for his “decisive role in the discovery of pulsars” — while Bell Burnell was overlooked. Bell Burnell herself has previously stated that she does not mind the oversight because she understands that Nobel prizes are not usually awarded to research students.

Her discovery still rings out half a century on, notes Mingarelli. For instance, earlier this year, NASA scientists demonstrated the feasibility of using pulsars to navigate, with their SEXTANT experiment on the International Space Station. The idea is that future robotic spacecraft could use the clockwork-like arrival times of X-rays emitted from pulsars to triangulate their locations. ■

INFORMATION

Google enters data ecosystem

Dataset Search could be especially helpful to cross-disciplinary researchers.

BY DAVIDE CASTELVECCHI

Google has unveiled a search engine to help researchers locate online data that are freely available for use. The company launched the service on 5 September, saying that it is aimed at “scientists, data journalists, data geeks, or anyone else”.

Dataset Search, now available alongside Google’s other specialized search engines, such as those for news and images — as well as Google Scholar and Google Books — locates files and databases according to how their owners have classified them. It does not read the content of the files themselves in the way search engines do for web pages.

Experts say that it fills a gap and could contribute significantly to the success of the open-data movement, which aims to make data freely available for use and re-use.

Government agencies, scientific publishers, research institutions and individual researchers maintain thousands of open-data repositories worldwide, containing millions of data sets.

But researchers who want to know what types of data are available, or who hope to locate data they know already exist, often have to rely on word of mouth, says Natasha Noy, a computer scientist at Google AI in Mountain View, California.

▶

▸ This problem is especially serious for early-career researchers who are not already “plugged” into a network of professional connections, Noy says. It’s also a downside for those who do cross-disciplinary research — for example, an epidemiologist who needs access to climate data that could be relevant to the spread of a virus.

CLASSIFIED SEARCH

Noy and her Google colleague Dan Brickley first described a strategy for solving that problem in a blogpost in January 2017 (see go.nature.com/2oounre).

Typical search engines work in two main stages. The first is to index the available pages by continuously trawling the Internet. The second is to rank those indexed pages, so that when a user enters search terms, the engine can provide results in order of relevance.

To aid search engines in indexing existing data sets, Noy and Brickley wrote, those who own the data sets should ‘tag’ them, using a standardized vocabulary called Schema.org, an initiative founded by Google and three other search-engine giants (Microsoft, Yahoo and Yandex), and which Brickley manages. The Google team also developed a special algorithm for ranking data sets in search results.

Given Google’s dominance in web searching, news that the company was moving into the data ecosystem quickly prompted major players to fall in line and standardize their metadata, says Mark Hahnel, chief executive of the data-sharing company Figshare in London. (Figshare is operated by the Holtzbrinck Publishing Group, which also has a majority share in *Nature*’s publisher.)

“By November, all the universities we’re working for had their stuff marked up,” Hahnel says. “I think this is a game changer for open data in the academia.”

Funding agencies sometimes mandate that research data be made available, and they are going to reach their ultimate goals only if the information is effectively recoverable, he says. “It legitimizes what the funders have been trying to do.”

AGENCY PARTNERSHIPS

An early supporter of Google’s experiment was the US National Oceanic and Atmospheric Administration (NOAA). The agency’s remit ranges from fisheries to the Sun’s corona, and its archives contain nearly 70,000 data sets — including ship logs from the 1800s. The trove adds up to more than 35 petabytes, comparable to the content of 35,000 typical hard drives.

Google’s tool will help NOAA to meet its open-data mission, says NOAA’s chief data officer Edward Kearns in Asheville, North Carolina. “We want to explore new ways to make those data available to others,” Kearns adds.

For Dataset Search to work, having the data owners’ collaboration was a crucial step. Although the system might become more sophisticated in the future, Google currently has no plans to actually read the data or analyse them, as it does with web pages or images. “A search tool like this one is only as good as the metadata that data publishers are willing to provide,” Noy says.

Like Google Scholar, Dataset Search currently offers no access for automated querying and no application programming interface (API) — although the company says that it might add that functionality in the future.

Noy says that as researchers begin to use Dataset Search, Google will watch how they interact with it and use that information to improve the search results. The company has no current plans to monetize the service, she says.

As Dataset Search evolves, it might also become integrated with Google Scholar, so that search results on a particular study could link to relevant data sets. ■

CORRECTION

The Editorial ‘What is Life?’ (*Nature* **561**, 6; 2018) misnamed Trinity College Dublin.

A Facebook data centre in Luleå, Sweden.

JONATHAN NACKSTRAND/AFP/GETTY

THE INFORMATION FACTORIES

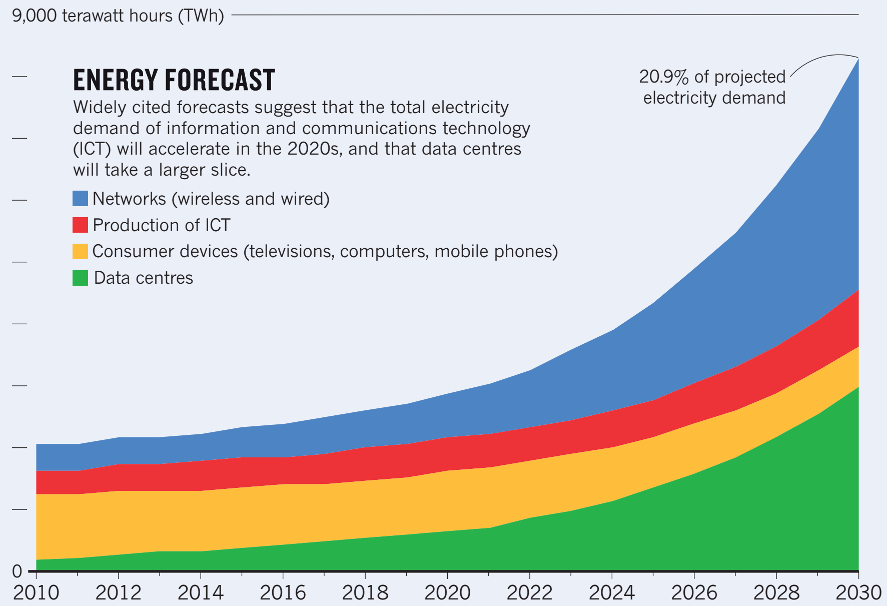
Data centres are chewing up vast amounts of energy — so researchers are trying to make them more efficient.

BY NICOLA JONES

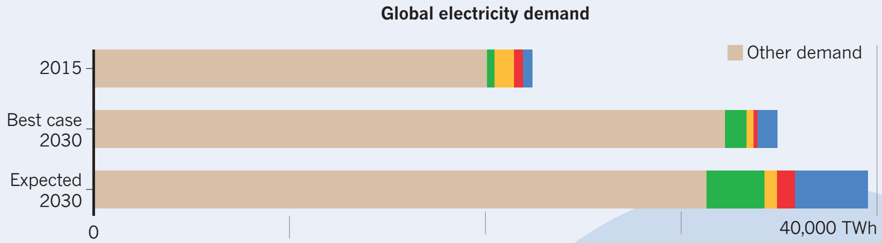
Upload your latest holiday photos to Facebook, and there's a chance they'll end up stored in Prineville, Oregon, a small town where the firm has built three giant data centres and is planning two more. Inside these vast factories, bigger than aircraft carriers, tens of thousands of circuit boards are racked row upon row, stretching down windowless halls so long that staff ride through the corridors on scooters.

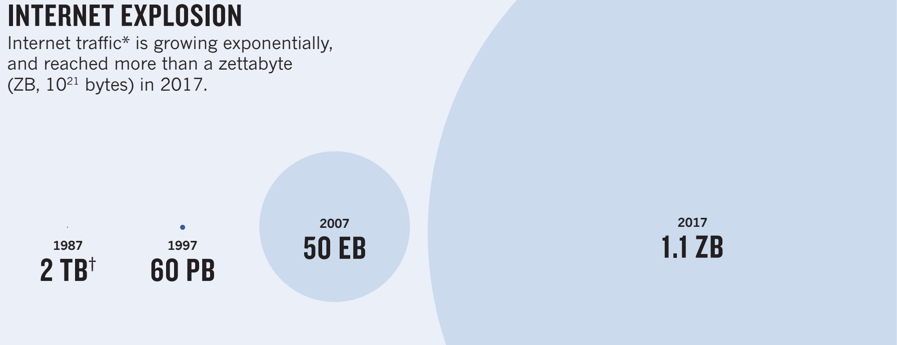
These huge buildings are the treasuries of the new industrial kings: the information traders. The five biggest global companies by market capitalization this year are currently Apple, Amazon, Alphabet, Microsoft and Facebook, replacing titans such as Shell and ExxonMobil. Although information factories might not spew out black smoke or grind greasy cogs, they are not bereft of environmental impact. As demand for Internet and mobile-phone traffic skyrockets, the information industry could lead to an explosion in energy use (see 'Energy forecast').

Already, data centres use an estimated 200 terawatt hours (TWh) each year. That is more than the national energy consumption of some countries, including Iran, but half of the electricity used for transport worldwide, and just 1% of global electricity demand (see 'Energy scale'). Data centres contribute around



The chart above is an 'expected case' projection from Anders Andrae, a specialist in sustainable ICT. In his 'best case' scenario, ICT grows to only 8% of total electricity demand by 2030, rather than to 21%.





*Traffic to and from data centres. †TB, terabyte (10^{12} bytes); PB, petabyte (10^{15} bytes); EB, exabyte (10^{18} bytes).

SOURCE: REF. 1

0.3% to overall carbon emissions, whereas the information and communications technology (ICT) ecosystem as a whole — under a sweeping definition that encompasses personal digital devices, mobile-phone networks and televisions — accounts for more than 2% of global emissions. That puts ICT's carbon footprint on a par with the aviation industry's emissions from fuel. What could happen in the future is hard to forecast. But one of the most worrying models predicts that electricity use by ICT could exceed 20% of the global total by the time a child born today reaches her teens, with data centres using more than one-third of that (see 'Energy forecast')[1]. If the computationally intensive cryptocurrency Bitcoin continues to grow, a sharp rise in energy demand could come sooner rather than later (see 'The Bitcoin bite').

For now, despite rising demand for data, ICT's electricity consumption is staying nearly flat, as increased Internet traffic and data loads are countered by increased efficiencies — including shuttering older facilities in favour of ultra-efficient centres such as Prineville's. But those easy wins could end within a decade. "The trend is good right now, but it's questionable what it's going to look like in 5–10 years," says Dale Sartor, who oversees the Center of Expertise for Energy Efficiency in Data Centers at the US Department of Energy's Lawrence Berkeley National Laboratory in Berkeley, California.

With the spectre of an energy-hungry future looming, scientists in academic labs and engineers at some of the world's wealthiest companies are exploring ways to keep the industry's environmental impact in check. They are streamlining computing processes, switching to renewables and investigating better ways to cool data centres and to recycle their waste heat. ICT's energy use must be "vigilantly managed", says Eric Masanet, an engineer at Northwestern University in Evanston, Illinois, who co-authored an International Energy Agency (IEA) report[2] last year on digitalization and energy — but if we stay on top of it, he says, we should keep future energy demand in check.

SHIFT TO HIGH GEAR

Perhaps the most startling forecast of ICT's future energy demand comes from Anders Andrae, who works on sustainable ICT at Huawei Technologies Sweden in Kista; he predicts that data-centre electricity use is likely to increase about 15-fold by 2030, to 8% of projected global demand[1]. Such dire numbers are controversial. "There have been many alarmist predictions of growing ICT energy use over the years, and all have proven to be bunk," says Masanet. Last year's IEA report estimated that although data-centre workloads will shoot up — tripling 2014 levels by 2020 — efficiency gains mean that their electricity demand might sneak up only by 3% (ref. 2). ICT's carbon footprint as a whole might even drift downwards by 2020, as smartphones take over from larger devices, researchers have suggested[3].

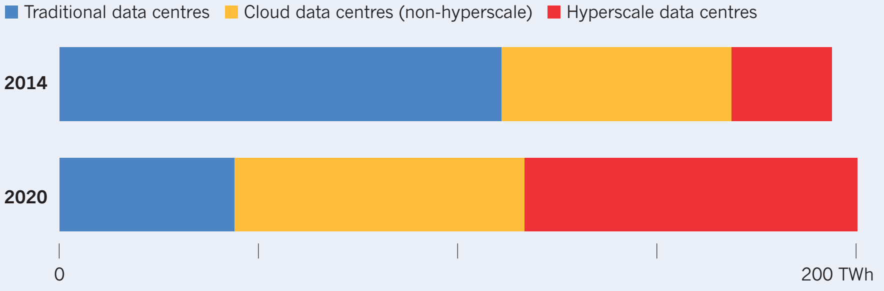
Data-centre electricity demand has remained roughly level over the past half-decade, in part because of the 'hyperscale shift' — the rise of super-efficient information factories that use an organized, uniform computing architecture that easily scales up to hundreds of thousands of servers. Hyperscale data centres emerged about a decade ago when companies such as Amazon and Google began to need fleets of a quarter of a million servers or more, says Bill Carter, chief technical officer at the Open Compute Project. It was started by Facebook in 2011 to share hardware and software solutions to make computing more energy-efficient. At that point, it made no sense to use off-the-shelf hardware from a computing firm, as companies had typically done.

"You had the opportunity to strip things down to just what you need, and make it specific to your application," Carter says. The new hyperscalers made bare-bones servers designed for purpose. "We stripped out video connectors, because there's no video monitor. There's no blinking lights because there's no one walking the racks. There's no screws," says Carter. On average, one server in a hyperscale centre is said to be able to replace 3.75 servers in a conventional centre.

The savings made by hyperscale centres can be seen in their power usage efficiency (PUE), defined as the total energy needed for everything, including lights and cooling, divided by the energy used for computing (a PUE of 1.0 would be a perfect score). Conventional data centres typically have a PUE of about 2.0; for hyperscale facilities, that's been whittled down to about 1.2. Google, for one, boasts a PUE of 1.12 on average for all its centres.

Older or less technologically adept data centres can contain a mix of equipment that is hard to optimize — and some that is even useless. In 2017, Jonathan Koomey, a California-based consultant and leading international expert on IT, surveyed with a colleague more than 16,000 servers tucked into corporate closets and basements and found that about one-quarter of them were "zombies", sucking up power without doing any useful work — perhaps because someone simply forgot to turn them off. "These are servers sitting around doing nothing except using electricity, and that's outrageous," says Koomey.

In a 2016 report, the Lawrence Berkeley National Laboratory estimated that if 80% of servers in small US data centres were moved over to hyperscale facilities, this would result in a 25% drop in energy use[4]. That move is under way. Today, the world has around 400 hyperscale data centres, many of them mopping up services for small corporations or universities that in the past would have had their own servers. Already they account for 20% of the world's data-centre electricity usage. By 2020, hyperscale centres will account for almost half of it, the IEA says (see 'Hyperscale shift').


HYPERSCALE SHIFT

Efficient 'hyperscale' data centres are predicted to swallow up half of data-centre electricity demand by 2020, as smaller, less-efficient centres shut down.

SOURCE: IEA

"WE'RE UP AGAINST THE LIMITS OF SHRINKAGE."

RUNNING HOT AND COLD

Once the hyperscalers have taken as much of the load as they can, further efficiencies will be harder to find. But corporations are trying. One emerging management technique is to make sure that servers are at full throttle as much of the time as possible, whereas others are turned off rather than being left idle. Facebook invented a system called Autoscale that reduces the number of servers that need to be on during low-traffic hours; in trials, this led to power savings of about 10–15%, the company reported in 2014.

One big way the hyperscalers have whittled down their PUE is by tackling cooling. In a conventional data centre, standard air conditioning can soak up 40% of the energy bill. The use of cooling towers, which evaporate water to drive the cooling of air, causes another environmental problem: US data centres are estimated to have used about 100 billion litres of water in 2014. Getting rid of compression chillers and cooling towers helps to save both energy and water.

One popular solution is to simply locate data centres in cool climes and blow the outside air into them. Such centres don't have to be in icy regions: Prineville is cool enough to take advantage of so-called 'free air cooling', and so are many other data-centre locations, says Ingmar Meijer, a physicist at IBM Research in Zurich, Switzerland.

Piped water is an even better conductor of heat, allowing centres to be cooled using warm water, which is less energy intensive to produce and recycle in the cooling system. Even in temperate climates, warm-water cooling has become the de facto solution for managing high-performance computers that are run fast and hot, including those in US Department of Energy labs and the Bavarian Academy of Sciences' SuperMUC supercomputer in Garching, Germany. Commercial centres in warm climates sometimes also invest in these systems, such as eBay's Project Mercury data centre in Phoenix, Arizona.

For high-density, high-power computing, the most efficient thing to do is to immerse servers in a non-conductive oil or mineral bath. Facebook trialled that in 2012 as a way to run its servers at higher speeds without overheating them. For now, immersion cooling is a specialist area with tricky maintenance, says Meijer.

In 2016, Google turned its DeepMind artificial intelligence (AI) research team to the task of tuning its data centre's cooling system to match the weather and other factors. Google says that, in tests, the team reduced its cooling energy bill by 40% and "produced the lowest PUE the site had ever seen". This August, the company announced it had turned cooling control at some data centres over to its AI algorithm.

Exploring innovative cooling solutions and making existing ones cheaper will become more important in coming years, says Carter. "As we connect the world, there's areas that won't be able to use free air cooling," he notes, pointing to Africa and southern Asia. And other developments will tax IT infrastructure in new ways. If self-driving cars flood onto the roads, for example, small server installations at the base of mobile-phone towers, used to help those cars communicate and process data, will need high-power devices that can handle real-time AI workloads — and better cooling options. This year, the Open Compute Project launched a project on advanced cooling, with the aim of making efficient cooling systems more accessible. "The hyperscalers have it figured out; they're extremely efficient," says Carter. "We're trying to help the other guys."

Hand-in-hand with better cooling is the idea of using the heat coming off the servers, thus

ENERGY SCALE

Global electricity demand
20,000 TWh

Electricity use by ICT
2,000 TWh

Data-centre electricity demand
200 TWh

Bitcoin use by mid-2018
20 TWh

Figures are approximate.

SOURCES: IEA/A. ANDRAE/REF. 6

saving electricity demand elsewhere. "It's like a free resource," says IBM researcher Patrick Ruch in Zurich. There are a few examples: the Condorcet data centre in Paris sends its waste heat directly into a neighbouring Climate Change Arboretum, where scientists study the impacts of high temperatures on vegetation. An IBM data centre in Switzerland warms a nearby swimming pool. But heat doesn't travel well, so waste-heat use tends to be limited to data centres located next to a convenient customer, or in a city that already uses piped hot water to heat homes.

A scattering of players are aiming to make waste heat more generally usable, including preliminary efforts to turn it into electricity. Others aim to use the waste heat to run cooling devices — IBM's US$2-million project THRIVE, for instance, is developing new materials that can better soak up water vapour and release it when exposed to heat, to make more efficient 'sorption heat pumps' to keep data centres cool.

POWER PLAY

At their core, data centres are only as good as the processors they're made of — and there's scope for improvement there, too. Since the 1940s, the number of operations a computer can perform with each kilowatt hour (KWh) of energy has doubled about once every 1.6 years for peak performance and every 2.6 years for average performance. That's a 10-billion-fold improvement over 50 years. By some measures, the rate of improvement has slowed since 2000, and the current generation of computing will run up against a physical barrier that limits the function of transistors in just a few decades, according to Koomey's calculations[5].

"We're up against the limits of shrinkage," says Koomey. Making comparable efficiency gains after that, he says, will require a revolution in how hardware is built and computing is done: perhaps through a switch to quantum computing. "It's basically impossible to predict," he says.

THE BITCOIN BITE

Cryptocurrency mining could push up energy demand

Since the cryptocurrency Bitcoin was born in 2008, concerns have grown that the energy demand of its production will escalate rapidly. Virtual coins are 'minted' by miners who buy specialized servers to crunch time-intensive computations in a growing blockchain that proves the validity of the new cryptocoins. By mid-2018, says Alex de Vries, a data consultant for international professional-services company PwC in Amsterdam, Bitcoin miners were probably using about 20 terawatt hours of electricity per year globally — less than 10% that of data centres, and less than 0.1% of total electricity use[6]. But estimates of how fast their usage is going up are contentious.

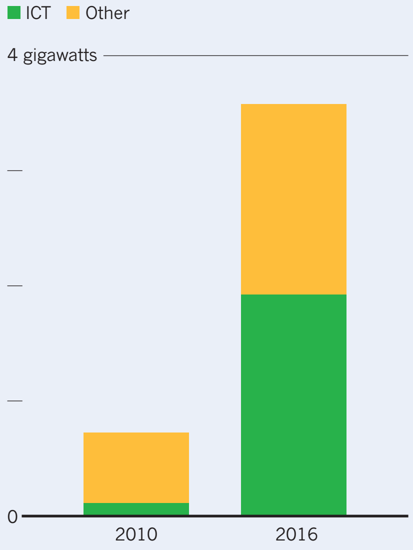
De Vries estimates that, by now, Bitcoin is gobbling up at least 0.33% of global electricity. Including other cryptocurrencies such as Ethereum bumps that up to 0.5%. "I think it's shocking," he says. But others, including Marc Bevand, a cryptocurrency researcher in San Diego, California, say that those numbers are inflated and based on gross assumptions. Bevand estimates that energy use might be as low as half of de Vries' current numbers by January 2019. "There is growth, but people are hyping this," says Jonathan Koomey, a California-based IT consultant, who is gathering data on cryptocurrency electricity consumption.

For now, Bitcoin mining is only profitable in places where electricity is cheap (about half the global average, says Bevand), which include China, Iceland and areas along the Columbia River in North America where hydroelectric power is plentiful. As Bitcoin miners dig in to an area and stress the grid, energy companies respond by raising their fees. That might prompt miners either to shut down or to take steps to drastically improve the energy efficiency of their hardware or system cooling.

Bitcoin could possibly be migrated to a less-energy-intensive blockchain system, says Bevand (as Ethereum is planning). Or, notes Koomey, "Say Bitcoin collapses for some reason; all those facilities will just go away." **N.J.**

GREEN GROWTH

The ICT industry dominates corporate agreements to purchase renewable electricity.



SOURCE: IEA

Although the core focus is on reducing ICT's energy usage, it's worth remembering that the information industry could also make our power use elsewhere smarter and more efficient. The IEA notes that if all vehicles become automated, for example, there is a utopian possibility that smoother traffic flow and easier carpooling would reduce the total energy demand of the transport industry by 60%. Buildings, which have accounted for 60% of the increased global electricity demand over the past 25 years, have huge scope for energy-efficiency improvements: smart heating and cooling, hooked up to building sensors and weather reports, could save 10% of their future energy demand. Chiara Venturini, director of the Global e-Sustainability Initiative, a Brussels-based industry association, estimates that the IT industry currently abates 1.5 times its own carbon footprint, and that could go up to almost 10 times by 2030.

ICT might also help to reduce global emissions by giving renewables a leg-up over fossil fuels. In 2010, the environmental group Greenpeace published its first ClickClean report, which ranked major companies and threw a spotlight on the environmental burden of IT. In 2011, Facebook made a commitment to using 100% renewable energy. Google and Apple followed in 2012. As of 2017, nearly 20 Internet companies had done the same. (Chinese Internet giants such as Baidu, Tencent and Alibaba, however, have not followed suit.) Back in 2010, IT companies were a negligible contributor to renewable-power purchase agreements with energy companies; by 2015, they accounted for more than half of such agreements (see 'Green growth'). Google is the largest corporate purchaser of renewable energy on the planet.

Cutting back our thirst for data might be the ultimate way to prevent energy use going into hyperdrive. But it's hard to see anyone agreeing to, say, limit their Netflix use, which accounts for more than one-third of Internet traffic in the United States. Banning high-definition colour cameras on phones alone could reduce data traffic in Europe by 40%, says Ian Bitterlin, a consulting engineer and data-centre expert in Cheltenham, UK. But, he adds, no one seems likely to dare to institute such rules. "We can't put the lid back on Pandora's box," he says. "But we could reduce data-centre power." ■

Nicola Jones *is a freelance journalist based in Pemberton, British Columbia, Canada.*

1. Andrae, A. & Edler, T. *Challenges* **6**, 117–157 (2015).
2. International Energy Agency. *Digitalization and Energy* (IEA, 2017).
3. Malmodin, J. & Lundén, D. *Sustainability* **10**, 3027 (2018).
4. Shehabi, A. *et al. United States Data Center Energy Usage Report* (LBNL, 2016); available at https://go.nature.com/ejg7sr
5. Koomey, J. G. *et al. IEEE Ann. Hist. Comp.* **33**, 46–54 (2011).
6. de Vries, A. *Joule* **2**, 801–805 (2018).

COMMENT

SPACE Astrophysics' long relationship with the military **p.173**

HISTORY At last, a definitive biography of Helmholtz **p.175**

PSYCHOLOGY How a questionable personality quiz went global **p.176**

ECOLOGY Pathogen could wipe out New Zealand's oldest tree species **p.177**

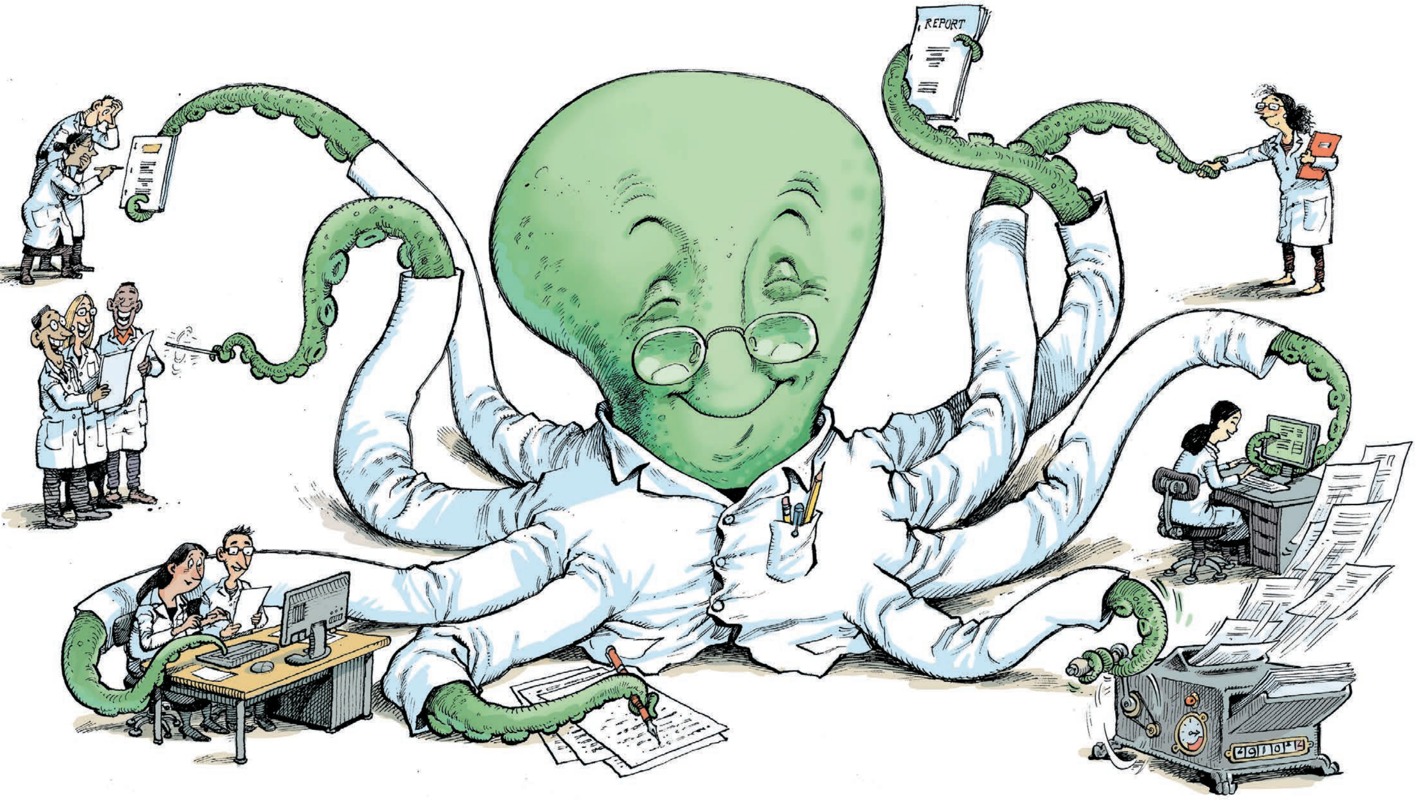


ILLUSTRATION BY DAVID PARKINS

The scientists who publish a paper every five days

To highlight uncertain norms in authorship, **John P. A. Ioannidis**, **Richard Klavans** and **Kevin W. Boyack** identified the most prolific scientists of recent years.

Authorship is the coin of scholarship — and some researchers are minting a lot. We searched Scopus for authors who had published more than 72 papers (the equivalent of one paper every 5 days) in any one calendar year between 2000 and 2016, a figure that many would consider implausibly prolific[1]. We found more than 9,000 individuals, and made every effort to count only 'full papers' — articles, conference papers, substantive comments and reviews — not editorials, letters to the editor and the like. We hoped that this could be a useful exercise in understanding what scientific authorship means.

We must be clear: we have no evidence that these authors are doing anything inappropriate. Some scientists who are members of large consortia could meet the criteria for authorship on a very high volume of papers. Our findings suggest that some fields or research teams have operationalized their own definitions of what authorship means.

The vast majority of hyperprolific authors (7,888 author records, 86%) published in physics. In high-energy and particle physics, projects are done by large international teams that can have upwards of 1,000 members. All participants are listed as authors as a mark of membership of the team, not for writing or revising the papers. We therefore excluded authors in physics.

Of what remained, 909 author records were Chinese or Korean names. Because Scopus disambiguates Chinese and Korean names imperfectly, these may have wrongly combined distinct individuals. For 2016 (when disambiguation had improved for Chinese and Korean names), at least ▶

12, and possibly more than 20, authors based in China were hyperprolific, the largest number from any country that year. We believe that this could be connected to Chinese policies that reward publication with cash or to possible corruption[2,3].

Because of the disambiguation issues, we excluded these names from further analysis, as well as group names and cases in which we found errors (such as journalistic news items misclassified as full articles), duplicate entries, or conference papers misassigned to an organizer.

This left 265 authors (see Supplementary Information). The number of hyperprolific authors (after our exclusions) grew about 20-fold between 2001 and 2014, and then levelled off (see 'Hyperprolific authors proliferate'). Over the same period, the total number of authors increased by 2.5-fold.

We e-mailed all 265 authors asking for their insights about how they reached this extremely productive class. The 81 replies are provided in the Supplementary Information. Common themes were: hard work; love of research; mentorship of very many young researchers; leadership of a research team, or even of many teams; extensive collaboration; working on multiple research areas or in core services; availability of suitable extensive resources and data; culmination of a large project; personal values such as generosity and sharing; experiences growing up; and sleeping only a few hours per day.

About half of the hyperprolific authors were in medical and life sciences (medicine $n = 101$, health sciences $n = 11$, brain $n = 17$, biology $n = 6$, infectious diseases $n = 3$). When we excluded conference papers, almost two-thirds belonged to medical and life sciences (86/131). Among the 265, 154 authors produced more than the equivalent of one paper every 5 days for 2 or more calendar years; 69 did so for 4 or more calendar years. Papers with 10–100 authors are common in these CVs, especially in medical and life sciences, but papers with the hundreds of authors seen in particle physics are uncommon.

Materials scientist Akihisa Inoue, former president of Tohoku University in Japan and a member of multiple prestigious academies, holds the record. He met our definition of being hyperprolific for 12 calendar years between 2000 and 2016. Since 1976, his name appears on 2,566 full papers indexed in Scopus. He has also retracted seven papers found to be self-duplications of previously published work[4]. We searched for news articles in Google detailing retractions for the next 20 most hyperprolific authors and found only one other author (Jeroen Bax) to have one retracted paper.

The 265 hyperprolific authors worked in 37 countries, with the highest number in the United States ($n = 50$), followed by Germany ($n = 28$) and Japan ($n = 27$). The proportion from the United States (19%) is roughly similar to its share of published science. Germany and Japan are over-represented. There were disproportionally more hyperprolific authors in Malaysia ($n = 13$) and Saudi Arabia ($n = 7$), countries both known to incentivize publication with cash rewards[5].

Hyperprolific authors also tended to cluster in particular institutions, often as part of a common study. For example, Erasmus University Rotterdam in the Netherlands had nine hyperprolific authors, more than any other institution. Seven of them co-authored mostly papers related to the Rotterdam study, a nearly 30-year-old epidemiological project, or its successor Generation R study, which have followed multiple health parameters in thousands of older adults and yielded thousands of publications. Five hyperprolific investigators from Harvard University in Cambridge, Massachusetts, also often co-authored papers related to cohort studies. Eleven hyperprolific authors across different institutions were on one large cohort study, the European Prospective Investigation on Cancer and Nutrition; other large epidemiological studies were also represented. Hyperprolific authors were also concentrated in cardiology and crystallography.

> **“Whether and how authorship is justified unavoidably varies for each author and each paper, and norms differ by field.”**

SURVEY

Criteria fulfilled?

One-third of the 81 authors identified as hyperprolific in 2016 replied when asked how often they met each of 4 criteria established for authorship of medical studies. Of the 27 responders, 19 admitted they had not met at least 1 criterion more than 25% of the time. Eleven wrote that they had not met two or more criteria upwards of 25% of the time.

- Substantial contributions to the conception or design of the work; or the acquisition, analysis or interpretation of the data for the work (9 of 27 met this criterion in less than 75% of their papers).
- Drafting the work or revising it critically for important intellectual content (9 of 27 met this criterion in less than 75% of their papers).
- Final approval of the version to be published (3 out of 27 met this criterion in less than 75% of their papers).
- Agreement to be accountable for all aspects of the work (14 out of 27 met this criterion in less than 75% of their papers).

These biological and medical disciplines with many hyperprolific authors exhibit different patterns from those found in particle and high-energy physics. Papers with hundreds to thousands of authors are the norm across a community of many thousands of scientists working in projects based at CERN, Europe's particle-physics laboratory near Geneva, Switzerland. In crystallography, papers tend to have few co-authors. In epidemiology and cardiology, long lists of authors appear only in relationship to specific research teams that seem to have a tradition of extensive authorship lists.

This raises the question of what authorship entails. The US National Institutes of Health, for example, has guidelines on the activities that qualify: actively supervising, designing and doing experiments, and data acquisition and analysis outside "very basic" work plus drafting the manuscript. Collecting funds or distant mentorship do not qualify. Most of the 6,000 authors in a recent survey across many geographical regions and disciplines felt that drafting a paper, interpreting results and analysing data should qualify for authorship, but attitudes varied by region and field[6].

AUTHORSHIP CRITERIA

Perhaps the most widely established requirements for authorship are the Vancouver criteria established by the International Committee of Medical Journal Editors in 1988. These specify that authors must do all of four things to qualify: play a part in designing or conducting experiments or processing results; help to write or revise the manuscript; approve the published version; and take responsibility for the article's contents.

The International Committee of Medical Journal Editors does not count supervision, mentoring or obtaining funding as sufficient for authorship. We did observe that some authors seemed to become hyperprolific on becoming full professors, department chairs or both. It is common and perhaps expected for scientists who assume leadership roles in large centres to accelerate their productivity. For example, clinical cardiologists publish more papers after they assume director roles (despite heavy clinical and administrative duties). Occasionally, the acceleration is stunning: at the peak of their productivity, some cardiologists publish 10 to 80 times more papers in one year compared with their average annual productivity when they were 35–42 years old. There was also often a sharp decrease after passing the chair to a successor. Another study noted similar patterns two decades ago[7].

One unexpected result was that some hyperprolific authors placed many publications in a single journal. Prominent in this regard were *Acta Crystallographica Section E: Structure Reports Online* (relaunched in 2014 as Section E: Crystallographic Communications, with brief structural data reports now published in IuCrData) and *Zeitschrift für Kristallographie New Crystal Structures*. Three authors have each published more than 600 articles in the former (Hoong-Kun Fun, Seik Weng Ng and Edward Tiekink); three authors have each published more than 400 papers in the latter (Karl Peters, Eva Maria Peters and Edward Tiekink). Three other authors (Anne Marie Api, Charlene Letizia, Sneha Bhatia) published many papers in single supplement issues of *Food and Chemistry Toxicology* focused on reviews of fragrance materials.

Journals indexed in Scopus are generally considered to be quality journals. The citation impact of hyperprolific authors was usually high, but there was large variability: with a median of 19,805 citations per author (range: 380 to 200,439). The median number of full papers per hyperprolific author in 2000–2016 was 677; across all hyperprolific authors, last author positions accounted for 42.5%, first author positions for 7.1%, and single authorships for 1.4%. Across the years, the median proportion of papers with middle author positions (that is, not a single, first or last author) was 51%, but varied from 2.1% to 98.5% for individual authors.

Our work to identify hyperprolific authors is admittedly crude. It is mainly intended to raise the larger question of what authorship entails. Whether and how authorship is justified unavoidably varies for each author and each paper, and norms differ by field. It is likely that sometimes authorship can be gamed, secured through coercion or provided as a favour. We could not assess these patterns in our data. We did not examine contributorship statements[8], which are not archived in Scopus. Nevertheless, even contributorship statements can be gamed and might not be accurate.

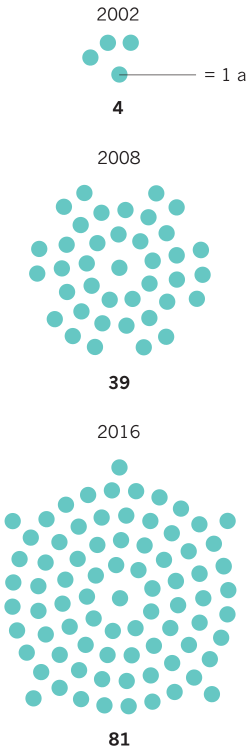
Further work is needed to explore how to best normalize these data and what is the optimal level of normalization: for example, adjusting for wide discipline, relatively narrow field and/or highly specific research team.

WHAT AUTHORS SAY

To better understand authorship norms, we e-mailed a survey to the 81 hyperprolific authors of 2016 (see Supplementary Information). We asked whether they fulfilled all four Vancouver criteria. Of the 27 who completed the survey, most said they did not (see 'Survey'). Almost all the responders were from US and European institutions. The only two responders from elsewhere stated that they failed Vancouver criteria in most of their papers. It is likely that the survey underestimates the proportion not meeting Vancouver criteria.

HYPERPROLIFIC AUTHORS PROLIFERATE

Numbers of authors with more than 72 papers a year increased dramatically over time.



SOURCE: J. P. A. IOANNIDIS, R. KLAVANS & K. W. BOYACK

Not all authors had approved the final versions of their own papers, but all considered approval of the final version necessary for authorship. Fifty-nine per cent (16 of 27) said that they had contributed more than any other listed author for 25 or more of the papers they authored in 2016.

Responses to the question "What, in your own words, do you think should be required for authorship?" generally reflected a requirement for "significant contributions", but also dissatisfaction with how authorship was assessed. One scientist said, "I personally don't count them as 'my papers' and don't have them on my CV as such, as there is a distinction between being a 'named author' versus a 'consortium member' authorship." Another observed that authorship was often awarded for seniority, and another that better distinctions were essential. "I think there should be levels of authorship — and not those implied by order!" It will be interesting to monitor how innovations in assigning credit, such as data citation or formal author contribution taxonomies, could alter authorship conventions. Authorship norms can vary within each field and even within each team. For example, some teams in epidemiology and cardiology apparently offer authorship more generously; others stick to stricter (and probably more appropriate) authorship criteria. For a similar task and contribution, one cohort study might credit 20 authors, another might give credit only to 3 people or none. For example, genome-wide studies typically include many dozens of authors. As a dramatic counter-example, one recent publication of a genome-wide study had only one author[9], and apparently that researcher did the same amount of work for which perhaps dozens would get authorship credit in similar papers spearheaded by different teams. Some evidence suggests that the increase in the average number of authors per paper does not reflect so much the genuine needs of team science as the pressure to 'publish or perish'[10].

Widely used citation and impact metrics should be adjusted accordingly. For instance, if adding more authors diminished the credit each author received, unwarranted multi-authorship might go down. We found that the 30 hyperprolific authors who seemed to benefit the most from co-authorship numbered 6 cardiologists and 24 epidemiologists (including those working on population genetics studies). (For these scientists, the ratio of their Hirsch H index to their co-authorship-adjusted Schreiber H_{m} index was higher; see Supplementary Information.)

Overall, hyperprolific authors might include some of the most energetic and excellent scientists. However, such modes of publishing might also reflect idiosyncratic field norms, to say the least. Loose definitions of authorship, and an unfortunate tendency to reduce assessments to counting papers, muddy how credit is assigned. One still needs to see the total publishing output of each scientist, benchmarked against norms for their field. And of course, there is no substitute for reading the papers and trying to understand what the authors have done. ■

John P. A. Ioannidis *is a professor of medicine at the Meta-Research Innovation Center at Stanford (METRICS), Stanford University, California, USA.* **Richard Klavans** *and* **Kevin W. Boyack** *are researchers at SciTech Strategies in Philadelphia, Pennsylvania, and Albuquerque, New Mexico, USA.*
e-mail: jioannid@stanford.edu

1. Wager, E., Singhvi, S. & Kleinert, S. *PeerJ* **3**, e1154 (2015).
2. Quan, W., Chen, B. & Shu, F. Preprint at https://arxiv.org/abs/1707.01162 (2017).
3. Hvistendahl, M. *Science* **342**, 1035–1039 (2013).
4. *Nature* **483**, 246 (2012).
5. Abritis, A., McCook, A. & Retraction Watch. *Science* **357**, 541 (2017).
6. Patience, G. S., Galli, F., Patience, P. A. & Boffito, D. C. Preprint at bioRxiv https://doi.org/10.1101/323519 (2018).
7. Drenth, J. *J. Am. Med. Assoc.* **280**, 219–221 (1998).
8. Sauermann, H, & Haeussler, C. *Sci. Adv.* **3**, e1700404 (2017).
9. Kim, S. K. *PLoS ONE* **13**, e0200785 (2018).
10. Papatheodorou, S. I., Trikalinos, T. A. & Ioannidis, J. P. *J. Clin. Epidemiol.* **61**, 546–551 (2008).

Supplementary Information accompanies this Comment online (see go.nature.com/2phadg2).

ROGER LEMOYNE/REDUX/EYEVINE

Young Rohingya refugees in Bangladesh in 2018.

Don't forget people in the use of big data for development

Information from satellites, mobile phones and other sources is not a panacea for international-development problems, warns **Joshua Blumenstock**.

Today, 95% of the global population has mobile-phone coverage, and the number of people who own a phone is rising fast (see 'Dialling up')[1]. Phones generate troves of personal data on billions of people, including those who live on a few dollars a day. So aid organizations, researchers and private companies are looking at ways in which this 'data revolution' could transform international development.

Some businesses are starting to make their data and tools available to those trying to solve humanitarian problems. The Earth-imaging company Planet in San Francisco, California, for example, makes its high-resolution satellite pictures freely available after natural disasters so that researchers and aid organizations can coordinate relief efforts. Meanwhile, organizations such as the World Bank and the United Nations are recruiting teams of data scientists to apply their skills in statistics and machine learning to challenges in international development.

But in the rush to find technological solutions to complex global problems there's a danger of researchers and others being distracted by the technology and losing track of the key hardships and constraints that are unique to each local context. Designing data-enabled applications that work in the real world will require a slower approach that pays much more attention to the people behind the numbers.

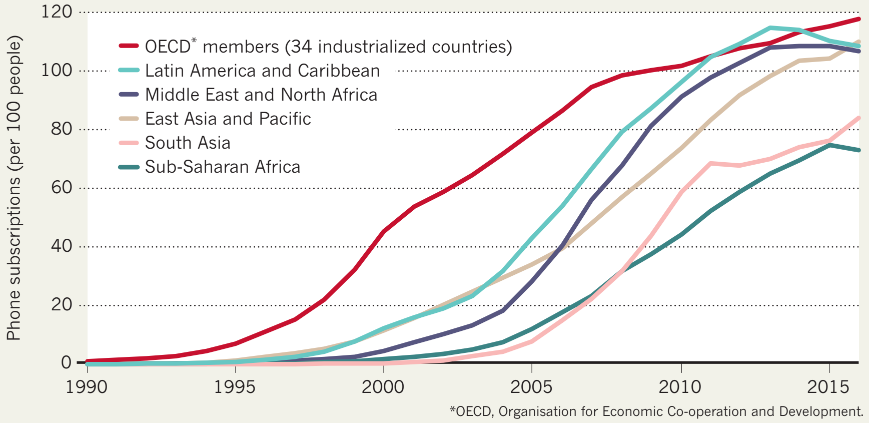
THE PROMISE

Already, data from mobile phones have transformed consumer lending in many developing countries[2]. Around five years ago, researchers discovered that some people, such as those who frequently make international calls or who have more Facebook friends than others in the same area, are more likely to repay their debts[3,4]. Machine-learning algorithms can detect these patterns and spit out credit scores for hundreds of millions of people who own a mobile phone, but who would otherwise be shut out of formal financial services because they lack collateral or access to a bank.

Other studies have shown that, with some tweaks, the same algorithms that Google, Facebook and other companies use to match advertisements to people online can also be used to match resources to people living in poverty[5,6]. These algorithms identify the 'digital signatures' of poverty in personal data from mobile-phone

DIALLING UP

The number of people owning mobile phones worldwide has soared over the past decade.



*OECD, Organisation for Economic Co-operation and Development.

SOURCE: INTL TELECOMMUN. UNION, WORLD TELECOMMUN./ICT DEV. REPORT & DATABASE

networks and in imagery from satellites. For instance, in most African countries, wealthier people tend to make more international calls than poorer people, who in turn are more likely than wealthier ones to live in houses with thatched roofs — as seen in satellite images. Studies from the past couple of years show that related approaches can be used to generate high-resolution maps of crop yields and childhood malnutrition[7,8].

In principle, such maps could enable governments and others to distribute humanitarian aid in a much more focused and timely way than they generally do now. Analysts estimate that the wealthy benefit more than the poor in one-quarter of all interventions aimed at reducing poverty. In two projects started in Armenia in 1996, only 8% of the tens of millions of dollars earmarked for the country's neediest citizens actually reached them[9].

Analysis of people's digital footprints could similarly improve public-health interventions during an epidemic, or assist national and international responses to crises. For instance, researchers have used phone data to reveal which neighbourhoods and individuals are most affected by natural disasters, where people relocate to, and how relocation affects the spread of disease[10–12]. Within a few years, it should be possible to track the effects of a natural disaster on individuals minute-by-minute — much as investors track the fluctuations of their stock portfolios.

THE PITFALLS

There are at least four problems with such tools.

Unanticipated effects. Solutions enabled by big data often bolster those who are already empowered rather than vulnerable people — largely because the power to derive value from the data tends to be concentrated in the hands of a few.

Take the example of 'digital credit'. Would-be borrowers are assessed using credit scores based on their history of phone use, and loans are dispatched instantaneously by mobile phone. A booming industry has developed since the first such service, M-Shwari, was launched in Kenya in 2012. Banks, phone companies and next-generation financial-service providers collectively make hundreds of thousands of loans per day in sub-Saharan Africa alone. Today, more than 25% of the Kenyan population has taken out at least one digital loan (see go.nature.com/2jytdp2).

So far, to my knowledge, no published study has documented whether these loans help people, or whether — like many short-term, high-interest 'payday' loans in the United States — they lead to poverty cycles and debt traps or prevent people from later obtaining loans from a bank as result of missing payments. Indeed, a substantial literature on microcredit, which predates digital credit by several decades, indicates that not everyone benefits from being able to borrow money[13].

Certainly, most digital-credit customers are first-time borrowers, and surveys suggest that many don't understand the terms of the loans they are being offered. For instance, a 2015 study in Rwanda found that only 51% of borrowers were aware of the interest rate they were being charged[14].

Risks of misappropriation extend beyond companies. Indeed, the potential for technology to be used in ways that don't necessarily benefit citizens could be much greater in countries where social institutions are weak and semi-authoritarian regimes more common. For example, various reports from China suggest that some people are being blocked from using trains and aeroplanes because of low 'social credit scores' — including those who have reportedly spread false information about terrorism or committed financial wrongdoings (see go.nature.com/2wwcnwq).

> ***"Solutions enabled by big data often bolster those who are already empowered."***

Lack of validation. Conventional data-collection methods in international development, which involve surveys and face-to-face interviews, are imperfect. But they have been developed over decades, and their limitations are well documented. By contrast, the flaws in the new approaches are not well understood. There is a risk that such tools will be deployed before they have been adequately tested.

With digital data, granular maps of the distribution of wealth in a nation can be produced at a fraction of the cost of a conventional household census. But the accuracy of such maps has been tested in only a handful of countries. And evidence suggests that patterns detected in one place do not always generalize. The tendency to make a lot of international phone calls correlates with wealth in Rwanda more strongly than it does in Afghanistan, for example[15].

More worrying is the lack of evidence that such algorithms will remain accurate over time. With colleagues, I have been working on interactive tools to provide real-time visualizations of population poverty and vulnerability. By benchmarking predictions with multiple rounds of survey data (including responses to questions about income, health and employment status), we've seen that the accuracy of our maps degrades quickly, sometimes within just a few months. Why might a model trained to predict wealth from phone data collected in the winter perform poorly in the summer? Because the relationship between poverty and phone use can change. For instance, wealthier people might make more international calls than poorer people during the holiday season, but that pattern could shift during the pilgrimage months, when many more people overall are travelling.

Finally, when people become aware of the fact that their personal data are being monitored to make decisions — for instance, about who gets humanitarian aid or who is eligible for a loan — they are inevitably incentivized to game the system. GiveDirectly, a non-profit organization in Africa and the United States that enables direct cash transfers to people living in poverty around the world, initially used satellite imagery to target aid to households with thatched roofs. But people soon caught on, to the point at which some would pretend to live in a thatched structure adjacent to their main iron-roofed house to become eligible for the aid.

Biased algorithms. When tools are trained on biased or patchy data, those who are poorly represented are often marginalized. This can be especially problematic for people in emerging economies: globally, the most disadvantaged people tend to be the least represented in new sources of digital data.

Representation can vary considerably even within such nations. Data from navigation apps such as Google Maps or Waze, for example, are increasingly being used to understand urban mobility[16]. But such apps typically require a smartphone, so any policy decisions made on this basis might primarily benefit the wealthier segments of society.

A mobile phone requires connectivity and electric power. Engaging with social media requires some literacy. And many digital-credit platforms require people to have a smartphone and a Facebook account. These prerequisites exclude vast segments of the population in developing countries.

Lack of regulation. Conventional development data are usually collected and disseminated by government agencies and aid organizations. The data that underpin artificial-intelligence applications are generally owned and controlled by private companies, who have little incentive to do anything except maximize profits.

In most wealthy countries, legislation is intended to limit governments' and companies' abuse of power. The US Supreme Court recently ruled that law-enforcement agencies cannot access phone data without a warrant. The General Data Protection Regulation in Europe is even more restrictive. In many developing nations, few such checks and balances exist, and those regulations that do exist are seldom enforced[17]. Currently, issues of data privacy, algorithmic transparency, fairness and accountability are off the radar of most companies operating in developing countries.

WAYS FORWARD

Several steps can be taken to try to address these concerns.

Validate. New sources of data should complement, not replace, old ones. Conventional data sets are essential to calibrate and validate big-data applications. And when tools such as poverty maps are used, they need to be benchmarked against existing methods.

An example of this two-pronged approach is ongoing work to evaluate the World Food Programme's aid efforts in Haiti. The organization is aware of the potential to save costs by using phone data. So, in collaboration with researchers, it is running a comparison, collecting phone and survey data side-by-side.

Customize. In most cases, the core technology being used is one that's been designed for a first-world purpose — say, automatically tagging Facebook photos with the names of friends. Finding that an algorithm can be adopted for a different use, such as identifying pockets of poverty in satellite photos, is a crucial insight — but further customization is needed before it can be useful to policymakers on the ground.

In the case of digital credit, for example, a learning algorithm might be remarkably accurate at predicting loan repayment, but the lending decision should also take local context into account. With this in mind, my colleagues and I are collaborating with Branch, a California company that provides micro-loans to millions in Africa, to research algorithms that weigh the borrower's default risk against the probable impact of a loan. The idea is to incorporate, from the start, a way to identify whether the loans are actually beneficial and to link each borrower to an 'impact score' as well as a 'credit score'. Insights emerging from the community of researchers focused on making machine-learning algorithms more fair, accountable and transparent (FAT) should help.

Deepen collaboration. Much of the innovation comes from the private sector — specifically from engineers in Silicon Valley in California. Many companies, including those pioneering digital credit, are motivated by the desire to do good as well as the promise of large profits. But development challenges cannot be tackled as '20%-time' side-projects. And next-generation solutions must be designed and produced by people who understand the problems and context — not just by those who understand the algorithms.

> *"New sources of data should complement, not replace, old sources."*

One way to achieve this is to foster collaboration between data scientists, development experts, governments, civil society and the private sector — and especially with people and organizations in the country in question. A step in the right direction is DataKind, a global network that attempts to link data scientists to social-change organizations, many of which are focused on issues in emerging economies. As are the Data for Development challenges: in 2012 and 2014, the Paris-based phone company Orange made troves of data available to researchers all around the world, which enabled early work on poverty mapping and urban planning. Meanwhile, fellowships, competitions, internships and perhaps some kind of year abroad offered to engineers in Silicon Valley could improve data scientists' understanding of the challenges people face in different countries.

Better still are efforts to increase technical capacity locally. Google and Facebook are funding an African Master's of Machine Intelligence: a one-year intensive programme that launches in Rwanda this month (see go.nature.com/2mcxjpc). Also encouraging are the three-week intensive summer boot camp now offered at the University of Cape Town in South Africa, and Data Science Africa, a conference conceived by African researchers. Likewise, the iHub, a creative work space in Nairobi, has helped to incubate hundreds of Kenyan start-ups. But these are exceptions, not the norm. Many more such efforts are needed.

A HUMBLER DATA SCIENCE

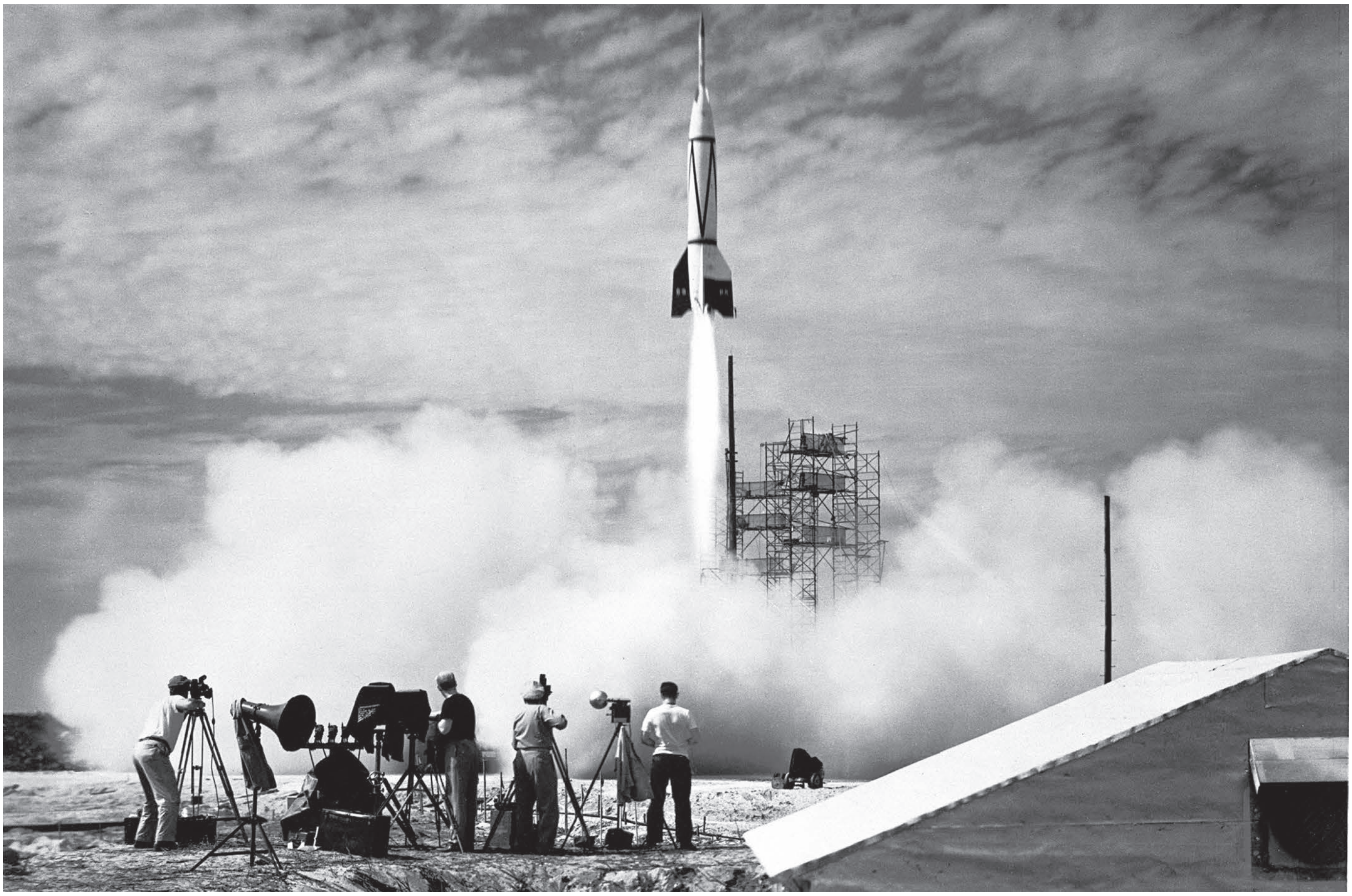
I am among those who are convinced that big data could transform international development. But many alleged silver bullets have missed their mark in recent decades. Think of the one laptop per child initiative. Hailed as a world-saver (see go.nature.com/2lgoaqh), the technology fizzled out because developers failed to understand the social and cultural environment in which it was rolled out[18].

Mike Driscoll, chief executive of the platform Metamarket, has described data science as "a blend of Red-Bull-fueled hacking and espresso-inspired statistics". In my view, the successful use of big data in development requires a version of data science that is considerably more humble than the one that has captured the popular imagination. ■

Joshua Blumenstock *is director of the Data-Intensive Development Lab and assistant professor at the School of Information, University of California, Berkeley, USA.*
e-mail: jblumenstock@berkeley.edu

1. GSM Association. *Unlocking Rural Coverage* (GSMA, 2017).
2. Francis, E., Blumenstock, J. & Robinson, J. BREAD Working Paper No. 516 (2017).
3. San Pedro, J., Proserpio, D. & Oliver, N. in *User Modeling, Adaptation and Personalization* (eds Ricci, F. *et al.*) 195–207 (Springer, 2015).
4. Björkegren, D. & Grissen, D. *Behavior Revealed in Mobile Phone Usage Predicts Loan Repayment* (SSRN, 2015).
5. Blumenstock, J., Cadamuro, G. & On, R. *Science* **350**, 1073–1076 (2015).
6. Jean, N. *et al. Science* **353**, 790–794 (2016).
7. Burke, M. & Lobell, D. B. *Proc. Natl Acad. Sci. USA* **114**, 2189–2194 (2017).
8. Osgood-Zimmerman, A. *et al. Nature* **555**, 41–47 (2018).
9. Coady, D., Grosh, M. & Hoddinott, J. *World Bank Res. Obser.* **19**, 61–85 (2004).
10. Lu, X., Bengtsson, L. & Holme, P. *Proc. Natl Acad. Sci. USA* **109**, 11576–11581 (2012).
11. Blumenstock, J. E., Eagle, N. & Fafchamps, M. *J. Dev. Econ.* **120**, 157–181 (2016).
12. Wesolowski, A. *et al. Proc. Natl Acad. Sci. USA* **112**, 11887–11892 (2015).
13. Banerjee, A., Karlan, D. & Zinman, J. *Am. Econ. J. Appl. Econ.* **7**, 1–21 (2015).
14. CGAP and InterMedia. *Financial Inclusion Insights: National Survey Report Rwanda* (CGAP/InterMedia, 2015).
15. Blumenstock, J. E. *AEA Pap. Proc.* **108**, 72–76 (2018).
16. Hanna, R., Kreindler, G. & Olken, B. A. *Science* **357**, 89–93 (2017).
17. Taylor, L. *Environ. Plan. D* **34**, 319–336 (2016).
18. Kraemer, K. L., Dedrick, J. & Sharma, P. *Commun. ACM* **52**, 66–73 (2009).

J.B. declares competing non-financial interests; see go.nature.com/2tzapmu.



UHA/UIG/BRIDGEMAN IMAGES

The United States tests a captured German V-2 rocket in 1946, technology that would later enable US astronauts to land on the Moon.

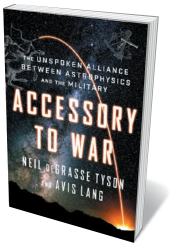
PHYSICS

Of arms and astrophysics

Sharon Weinberger applauds a study of the fraught nexus of science and war.

A poignant moment of professional soul-searching sets the scene in US astrophysicist Neil deGrasse Tyson's latest book, *Accessory to War*. In 2003, Tyson was attending the 19th annual Space Foundation conference in Colorado Springs, Colorado — a gathering that drew together a mix of scientific, industrial and military personnel with a common interest in the cosmos. Up to that point, he had believed that the conference was focused on the peaceful study of space, even if it involved "a bit of arms trading on the side". But that year, the gathering coincided with the opening volleys of the US-led invasion of Iraq. Tyson suddenly recognized that the conference represented a greater reality: the collision and collusion of civil science and military space.

He investigates that relationship between the military and astrophysicists — along with space scientists, aerospace engineers and others in broadly related fields — in depth in the book, co-authored by writer-researcher Avis Lang. Tyson is well placed to explore this territory, as an eminent scientist and science popularizer who has also served on US government commissions, including one aimed at strengthening the aerospace industry.



Accessory to War: The Unspoken Alliance Between Astrophysics and the Military
NEIL DEGRASSE TYSON & AVIS LANG
W. W. Norton (2018)

Accessory to War traces an epic sweep of the scientific study of the cosmos, starting with the proto-astronomers of ancient Mesopotamia and proceeding through centuries of discovery and technological advance. With brutal honesty, Tyson and Lang reveal how looking at the stars has always been both 'pure' science and an adjunct of war-making. From star charts and chronographs to telescopes and satellites, the military has made use of astronomical instruments and inventions for hundreds of years; in turn, astrophysicists gained funding and strategic support. As they write, "each side is accessory to the other's needs, interests, and resources".

AGE OF DISCOVERY

This pattern became prominent in early modern Europe, when hunger for power, discovery and expansion fuelled the development of astrophysics. In the fifteenth and sixteenth centuries, the emerging field of astronomy became a tool of empire, as nations used their growing knowledge of the cosmos and instrumentation such as astrolabes to help them to navigate and explore the seas. The 'voyages of discovery' that began with Christopher Columbus and Ferdinand Magellan fuelled knowledge,

▸ but they also involved violence and privateering: by the seventeenth century, huge ships "laden with cargo and cannon" filled sea lanes. And, as Tyson and Lang write, it was "the hegemonistic agenda of the British empire builders" that drove the 1769 expedition to observe the transit of Venus by Captain James Cook and naturalist Joseph Banks, among others.

It wasn't just lust for power that helped to make astrophysics "a historical handmaiden" of war, they argue. The relationship was deepened by the cost of the equipment, which made it dependent on national coffers; even Galileo Galilei sought patronage for his 'spyglass' by recommending its usefulness in battle. That collaboration continued over the centuries. In the heat of the Second World War, Nazi Germany's military paid to develop the V-2 rockets first launched at Britain and other targets in 1944; more than two decades later, the same technology helped US astronauts to land on the Moon. And the Earth-imaging satellites currently used to plot deforestation and understand climate change were first launched in 1959 as part of the covert US Corona programme to keep tabs on the Soviet Union.

COMPETITIVE DRIVE

Tyson is willing to challenge some of the long-standing tenets of astrophysics, notably the presumption that the quest for knowledge should drive the peaceful exploration of science. The reality is different. Not only has military competition led to space exploration, but the lack of a military imperative can doom such efforts. "Faced with a Cold War space adversary, the United States placed the bootprints of twelve astronauts on the dusty lunar surface," Tyson and Lang write. In peacetime, it sends astronauts to "boldly go where hundreds have gone before": the International Space Station in low-Earth orbit. Perhaps this side benefit of competition is even why Tyson has spoken with some approval of US President Donald Trump's proposed new branch of the military, the Space Force.

The authors go so far as to argue that astrophysicists are "hidden casualties of peace". The cold war's end did not herald a golden era of halcyon scientific pursuits. Instead, it resulted in the cancellation of major US projects such as the Superconducting Super Collider, the proposed particle accelerator near Waxahachie, Texas, that was abandoned in 1993. Put like that, might astrophysicists hope for war, at least if it came with a nifty new science facility?

ANN RONAN PICTURES/GETTY

Werner von Braun (right) perfected V-2 rocket technology during the Second World War.

IBM 7090
DATA PROCESSING SYSTEM

Galileo Galilei's telescope.

FINE ART IMAGES/GETTY

CONSPICUOUS SILENCE

Accessory to War is thought-provoking and reflective, but does have one odd omission: the voices of today's 'handmaidens' are largely missing. By not giving voice to more scientists, particularly those working with the military, Tyson and Lang miss some subtleties in the interactions between researchers and war-makers. For example, they describe the history of the Arecibo radio telescope in Puerto Rico, funded by the Pentagon's technical innovation agency — now called the Defense Advanced Research Projects Agency — ostensibly to advance missile defence. That instrument, they write, could be used to measure the radar signatures of warheads, to help missile-defence systems shoot them down. "Oh, and the telescope could do astrophysics on the side," write Tyson and Lang. Actually, it was the other way around. The Pentagon scientists most closely associated with Arecibo have long acknowledged that it had no real use for missile defence; this indicates that its alleged military utility was nominal justification for a facility that the reseachers thought would be good for science. In effect, the military research took second place. The scientists had snookered their military masters.

Even more importantly, the lack of perspective from astrophysicists today leaves the reader with a key question. What moral conflicts do scientists face, and how do they justify their work with the military? Do researchers feel any qualms about using the Hubble Space Telescope, knowing it was developed as part of a spy satellite system? The only insight we get into the psyche of an astrophysicist is Tyson's ethical dilemma on attending that 2003 space conference as bombs fell on Baghdad.

However horrified, Tyson stayed at the conference. Without war, he writes, there would be "no astronomy, no astrophysics, no astronauts, no exploration of the solar system, and barely any comprehension of the cosmos". From him we learn that the handmaiden's bond to her mistress, whatever her flaws, is too strong to break — even for a morally outraged astrophysicist. ■

Sharon Weinberger *is the author of* The Imagineers of War: The Untold Story of DARPA, the Pentagon Agency That Changed the World.
e-mail: sharonweinberger@gmail.com

SCIENCE HISTORY

The last polymath

Henning Schmidgen praises a tome on Helmholtz, titan of nineteenth-century science.

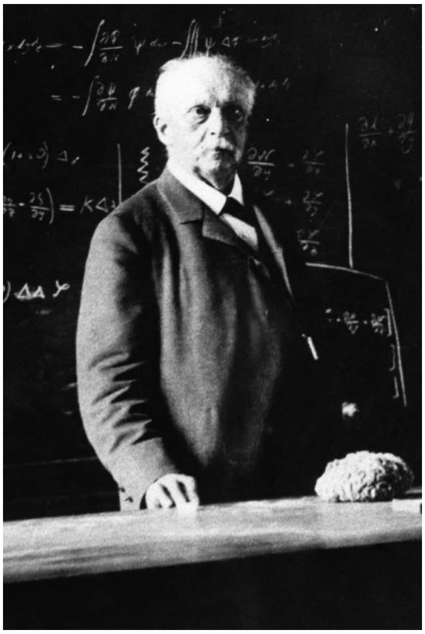
Hermann von Helmholtz is a towering figure in the history of modern science. In the nineteenth century, his peers in experimental physiology and physics regarded him as a scientific genius, and the German public nicknamed him 'Imperial Chancellor of the Sciences', given his multifaceted activities as a manager of scientific institutions. These ranged from his own laboratories in Heidelberg and Berlin to the Imperial Physical Technical Institute, established in Berlin with financial support from a friend, industrialist Werner von Siemens. Moreover, in contrast to Charles Darwin or Louis Pasteur (or, later, Albert Einstein), Helmholtz's pathbreaking achievements were not confined to one discipline.

Against this background, historians and philosophers consider Helmholtz to be one of the last — if not the last — polymaths in the history of science. His is a legacy so rich it is difficult to grasp.

Born in 1821, Helmholtz trained as a physician at the University of Berlin. In the 1840s and 1850s, he devised the ophthalmoscope — an instrument for examining the inside of the eye — and took innovative measurements of the speed of nerve impulses. He performed extensive experiments in sensory physiology, resulting in multi-volume manuals on optics and acoustics, published during the 1860s. He then established himself as a physicist. During the 1870s and 1880s, his main fields of research were electrodynamics and thermodynamics. By the time he died, in 1894, Helmholtz had gained further recognition as a philosopher of science and nature in the mould of Immanuel Kant, advocating precision work in the laboratory alongside sophisticated theory.

Historian David Cahan copes confidently with these complexities in his monumental new biography. The result of almost three decades of scholarly work, *Helmholtz* is a comprehensive and timely account. In recent years, historians of science have published several studies devoted to isolated aspects of Helmholtz's work — his contributions to neurophysiology and hydrodynamics, as well as his epistemology and aesthetics.

However, anyone interested in the complete picture was forced to go back more than 100 years, to a bulky biography by German mathematician Leo Königsberger. In no fewer than three volumes published in 1902–03, Königsberger, a friend of the Helmholtz family, depicted the German scholar as a lonesome genius whose scientific success was the result of unusual giftedness combined with extremely hard work.



EMILIO SEGRE VISUAL ARCHIVES/AIP/SPL

Hermann von Helmholtz in 1894.

Helmholtz is an impressive corrective to such partial or simplistic treatments. The book not only accounts for the German scholar's voluminous publications (his *Lectures on Theoretical Physics* alone comprise six volumes, posthumously published between 1897 and 1907), but also encompasses an uncounted number of manuscripts and letters dispersed in archives all over the world. The result is as compelling as it is convincing.

Cahan's Helmholtz emerges as a liberal conservative devoted to his family, and as a talented and ambitious scientist whose career unfolded in close connection with the culture and society of Imperial Germany. This is not to downplay the significance of Helmholtz's achievements.

On the contrary, Cahan aims to illuminate these achievements — their texture, as it were. He ranges over Helmholtz's work: the introduction of innovative instruments and experimental techniques; extensive writing, often aimed at securing priority for his discoveries and inventions; and intense communication with scientists, engineers and politicians in Germany and abroad. Helmholtz, we learn, was a skilled networker and salesperson of all things scientific, as well as an exceptional researcher.

Drawing on recent approaches to the history of science that highlight cultural and social context, *Helmholtz* reveals the preconditions for the successful reception and propagation of Helmholtz's work. It becomes clear that, in the late 1840s, Helmholtz referred to and relied on German telegraph technology when performing his precision measurements of nerve impulses. Cahan also shows that his organizational achievements as 'Imperial Chancellor of the Sciences' would not have been possible without the political and administrative network that he and his colleague Emil du Bois-Reymond began to weave at the start of their careers, in Berlin. Readers of Cahan's wonderful book will even be convinced that Helmholtz's laboratory work on physiological optics and acoustics became popular among artists, because he met painters such as Adolph von Menzel, and musical luminaries including Richard Wagner.

Helmholtz: A Life in Science
DAVID CAHAN
University of Chicago Press (2018)

Cahan effortlessly depicts and analyses Helmholtz's trajectory as a movement through various disciplinary, institutional and cultural contexts. His account is not a simple biography or a mere retracing of events. As the subtitle indicates, its ambitious goal is to explore the relationships between life and science.

The guiding thread of this exploration is what Cahan calls the "civilizing power" of science. For Helmholtz, science was neither blunt profession nor pure vocation; it was a way of life. Truth was the ultimate goal, reached through the crucial means and methods of empirical care, precision and the reproducibility of findings. Science in the Helmholtzian sense is vitally important not just for individuals, but also for the society that they belong to.

Cahan demonstrates that Helmholtz strikingly embodies the cognitive and ethical potential of science. His book is the new standard for all Helmholtz scholars, a comprehensive guide and crucial resource for historians, philosophers or sociologists interested in this scientific luminary. *Helmholtz* is also a highly topical book for those impressed by and convinced of the social significance of science. ■

Henning Schmidgen *is a historian of science and professor of media studies at Bauhaus University in Weimar, Germany. He is the author of* The Helmholtz Curves.
e-mail: henning.schmidgen@uni-weimar.de

PSYCHOLOGY

Self-made creators of Myers–Briggs

S. Alexander Haslam enjoys the tale of how a questionable personality quiz went global.

Most of us have, at some point, completed a formal personality test. Chances are it was the Myers-Briggs Type Indicator (MBTI). Introduced by Isabel Briggs Myers in 1943, it asks 93 questions, such as, "If you were a teacher, would you rather teach (a) Fact courses or (b) Courses involving theory?" On the basis of their answers, an individual's 'personality type' is plotted along four axes: extraversion (E) to introversion (I); sensing (S) to intuition (N); thinking (T) to feeling (F); and judging (J) to perceiving (P). The results, proponents claim, allow people to understand themselves and others in terms of a four-letter label: I may be an ESTJ (a 'supervisor type'); you an INTP (an 'architect type').

By conventional psychometric criteria, the MBTI is highly suspect. In failing to measure what it purports to measure, it has low validity and in failing to elicit consistent responses across testing contexts, it has low reliability and little predictive power. But *The Personality Brokers*, by literature scholar Merve Emre, does not debate the limitations of the MBTI — a path many have trodden. Instead, Emre's careful investigations of the tool's bizarre origins and alarming impact weave a compelling narrative that recounts the rise of twentieth-century managerial and personnel-theory science with the gritty wistfulness of a John Steinbeck novel.

Isabel Briggs Myers (1897–1980) was an autodidact who eschewed formal psychological methods of test development and validation. She became interested in personology, as she called it, largely as a result of an obsession her mother, Katharine Cook Briggs (1875–1968), had with the ideas of psychoanalyst Carl Jung. Emre charts the women's competitive relationship and expanding ambitions with sensitivity and skill.

Initially Cook Briggs wanted to make a landmark contribution to the practice of child-rearing. In popular-magazine articles, she presented Isabel as a triumph of an "obedience-creativity" regime. In this model, kindness, warmth and play were won only after authoritarian orders to study and work had been complied with. Before reading Briggs Myers bedtime stories, Cook Briggs required her to complete a demanding programme of study. By her early 30s, Briggs Myers was an accomplished polymath and award-winning writer of formulaic but engaging detective fiction.

These ventures paved the way for both women's fervent interest in personality. Cook Briggs hoped to make Jung's obscure writings accessible to the world by cataloguing the character of everyone she met on index cards. Briggs Myers formalized her mother's project into the MBTI, after losing the proceeds of her novels in the 1930s economic crash.

Isabel Briggs Myers (left) and her mother, Katherine Cook Briggs (right).

MYERS & BRIGGS FOUNDATION

A breakthrough came at the end of the war, when the test was embraced by the Office of Strategic Services (a short-lived forerunner to the Central Intelligence Agency) as a tool for assigning intelligence operatives in newly liberated Europe. Briggs Myers recognized and promoted the tool's attractiveness to those responsible for sorting students into suitable fields and employees into jobs. By the mid-1950s, she had persuaded more than 45 medical schools to use it for admissions and clients such as General Electric, Standard Oil and Bell Laboratories to inform recruitment and promotion decisions. Today, the MBTI is deployed widely, although its uses remain largely corporate: building a team or recruiting for an executive-training programme.

> *"This is a tool of status quo, not social change."*

Emre's account of the emergence and appeal of the MBTI is nuanced and non-judgemental. She imparts fascinating details of two willful women making their way in a sexist — and racist — corporate landscape (observing that the MBTI's questions spoke to a middle-class world that was then largely off-limits to people of colour). She excels in showing how the MBTI's success rests on its facile summary of a seemingly unfathomable subject: the self. Moreover, because it produces profiles that are always positive, it proves perenially palatable to consumers.

The Personality Brokers: The Strange History of Myers-Briggs and the Birth of Personality Testing
MERVE EMRE
Doubleday (2018)

Hence, as Emre explains, the MBTI also emerged as the perfect device for the creation of "The Organization Man", as heralded by William H. Whyte in his 1956 best-selling text of the same name. Specifically, the test required employees to submit to the rigorous regimes of corporate life, and fetishized the need for them to be "individuals". This fed into hire-and-fire managerial prerogatives while also extinguishing the appetite for collective resistance. For all its light and liberal pretensions, the MBTI thus has a dark and deeply conservative heart. This is a tool of status quo, not social change. Indeed, its political functions are plain to see, precisely because the research that underpins it is so poor.

Emre's engagement with some matters of psychology is less assured. She might have done more to help readers understand how the Myers-Briggs instrument and industry relates to, and departs from, theories and methodologies of contemporary psychological science. She should have avoided name-dropping psychologists, such as Stanley Milgram and Philip Zimbardo, whose work is neither relevant nor helpful to her cause.

And, notwithstanding the blurb's claims, if you want to know "what makes you you", look elsewhere — not least, to discover up-to-date research pointing to the dynamic nature of personality and to its collective, rather than purely personal dimensions. Indeed, using the MBTI to unlock the mysteries of the self is akin to trying to understand the Stone Age by watching *The Flintstones*.

But Emre's goal is not to understand the self. Rather, it is to explore two particular selves. This is a story of two remarkable women who founded an enormously successful commercial empire, which has significantly framed the conversation about the contours of personality, as understood by millions of Westerners today. As Emre shows convincingly, the fact that the premises of this conversation are so flimsy makes the examination of its origins and impact all the more interesting, more important — and more disturbing. ■

S. Alexander Haslam *is a professor of psychology at the University of Queensland in Brisbane, Australia.*
e-mail: a.haslam@uq.edu.au

Correspondence

DANNI THOMPSON/ALAMY

Save Maori people's sacred tree species

New Zealand's largest and oldest living tree species, the kauri (*Agathis australis*), is threatened with extinction by an invasive pathogen, *Phytophthora agathidicida*. These conifers live for more than 2,000 years, reaching heights in excess of 50 metres and girths of up to 14 metres. They are cultural icons, regarded as living ancestors by the indigenous Maori people. With no known treatment against kauri dieback, we stress that the only immediate option to halt the soil-borne pathogen's spread is complete closure of kauri forests.

This drastic proposal has met with strong resistance because of the economic implications of losing a million or so visitors every year. The collective action from mandated government agencies has so far been inconsistent and ineffective, despite reasonable resources. Indigenous and community groups are being left to develop their own management plans in isolation.

As scientists, we know that the losses we face are too catastrophic to risk piecemeal measures. The government must push for urgent research into controlling the pathogen. Until that is better understood, all kauri forests should remain shut.

Amanda Black, Nick Waipara *Lincoln University, Lincoln, New Zealand.*
Monica Gerth *Victoria University, Wellington, New Zealand.*
amanda.black@lincoln.ac.nz

A hiker looks up at New Zealand's largest giant kauri tree (*Agathis australis*).

Tighten attribution of weather events

Your claim that the connection between extreme weather and climate change is "now routine and reliable science" misrepresents the current state of climate-extreme attribution science (see *Nature* **560**, 5; 2018).

Although methods to quantify the contribution of climate change to particular extreme weather events are developing rapidly, the science is at an early stage, and there is as yet no consensus on which approach is best. We are still a long way from achieving high confidence in quantitative results (see go.nature.com/2wlyuc6).

Attribution depends fundamentally on global climate models that can adequately capture regional weather phenomena — including circulation anomalies such as the weak jet stream and large, persistent planetary-scale atmospheric waves that characterized this summer's weather. Accurate simulation of such extremes remains a challenge for today's models.

It is not enough to increase the size of the ensemble of simulations if the models themselves have fundamental limitations. Any statement on attribution should therefore always be accompanied by a scientifically robust demonstration of the model's ability to simulate the global and regional weather patterns and the related weather phenomena that lie at the root of extreme events.

Rowan Sutton* *National Centre for Atmospheric Science, University of Reading, UK.*
**On behalf of 5 co-signatories (see go.nature.com/2n4zys3).*
rowan.sutton@ncas.ac.uk

Undergrad research: classroom training

I have mentored seven undergraduate research students over the past three years, all of whom have been co-authors on scientific papers and conference abstracts. In my view, their success had less to do with their academic prowess or even my mentoring abilities, and more to do with their preparation before engaging in research (see also J. Ankrum *Nature* http://doi.org/gdwps2; 2018; and J. Trant *Nature* **560**, 307; 2018).

These students had previously taken courses that were largely enquiry-based. They participated in multi-part projects or case studies — for example, collaborating over several class periods to address a specific problem. They were given real data sets to sharpen their analytical skills. In these projects, formidable obstacles were put in their way to stimulate awareness of thought processes (metacognition), a useful faculty when it comes to dealing with uncertainty and unpredictability in research results.

This classroom training in teamwork, critical analysis of the literature and reasoning like researchers helps to equip undergraduate students for the research lab, so that they can start contributing to its productivity from the outset.

Andrew A. David *Clarkson University, Potsdam, New York, USA.*
adavid@clarkson.edu

Teach PhD students to peer-review

I suspect that journal editors could be reluctant to seek out early-career scientists as peer reviewers (see M. Casado *Nature* **560**, 307; 2018). Researchers at this stage are often invisible to editors, because they have no publications or online research profile. Senior scientists, who have the grounding and scope necessary for in-depth peer review, are seen as a safer bet. In my experience, however, training graduate students in the critical assessment of research papers goes a long way towards equipping them for the task.

I ask graduate students participating in a research-methods course to independently review the same selection of manuscripts before these are submitted for publication. They then collectively discuss the merits or inadequacies of each paper. In the process, the students learn how peer review works, how to avoid pitfalls in publishing and how to prepare and improve their own papers.

Ultimately, such group peer-teaching exercises could help to improve the peer-review process for all researchers, even the experienced ones.

Tony R. Walker *Dalhousie University, Halifax, Nova Scotia, Canada.*
trwalker@dal.ca

PHYSIOLOGY

Receptor becomes a ligand

The protein RANKL is released by bone-forming cells called osteoblasts, and binds to its receptor, RANK, on osteoclast cells to trigger bone removal. It emerges that the pathway can act in reverse to stimulate bone formation. See Article p.195

MONE ZAIDI & CHRISTOPHER P. CARDOZO

The integrity of the adult skeleton must be maintained by tightly coupled bone-remodelling processes: old bone is resorbed by cells called osteoclasts and is replaced by new bone, which is synthesized by cells called osteoblasts[1]. Osteoblasts become buried within the newly formed bone and morph into a third cell type, osteocytes. Osteoblasts and osteocytes secrete the signalling protein RANKL, which activates the RANK receptor on haematopoietic stem cells, triggering their differentiation into osteoclasts and so promoting osteoclast function[2,3]. But it is less clear how osteoclasts signal to osteoblasts to modify bone formation. On page 195, Ikebuchi *et al.*[4] provide compelling evidence that osteoclasts regulate osteoblasts using the same RANKL–RANK system acting in reverse — in this setting, it is the RANKL ligand that modulates intracellular signalling, rather than the RANK receptor.

The authors first confirmed previous data[5] indicating that osteoclasts release small extracellular vesicles that harbour RANK on their surface (Fig. 1a). They showed that overlaying a layer of isolated RANK-bearing vesicles on a layer of mouse osteoblasts in culture activated the expression of the differentiation-promoting genes *Col1a1, Runx2* and *Osx* in the cells. The presence of RANK-containing vesicles also triggered mineral deposition by osteoblasts, both *in vitro* and in mouse skulls *in vivo*. Mineral deposition, indicative of bone formation, was promoted by the Runx2 protein.

Runx2 is known[6] to be activated by a signalling pathway involving the proteins PI3K, Akt and mTOR. Indeed, Ikebuchi *et al.* found that Runx2 activation in mouse osteoblasts was blocked by the mTOR-inhibitor molecule rapamycin. The authors then showed that RANKL activates the PI3K–Akt–mTOR pathway through a region in its cytoplasmic tail that is rich in the amino acid proline; this proline-rich domain interacts with kinase enzymes that activate PI3K. Mutations in crucial proline residues within this domain abrogated the effects of RANK–RANKL reverse signalling.

Finally, Ikebuchi and colleagues established that RANKL monomers are not activated by RANK. Instead, crosslinking of individual RANKL proteins to produce multimers is a prerequisite for downstream signalling. The authors showed that genetically modified anti-RANKL antibodies containing structures called leucine zippers, which are known to induce trimer formation, could activate RANKL reverse signalling in osteoblasts (Fig. 1b).

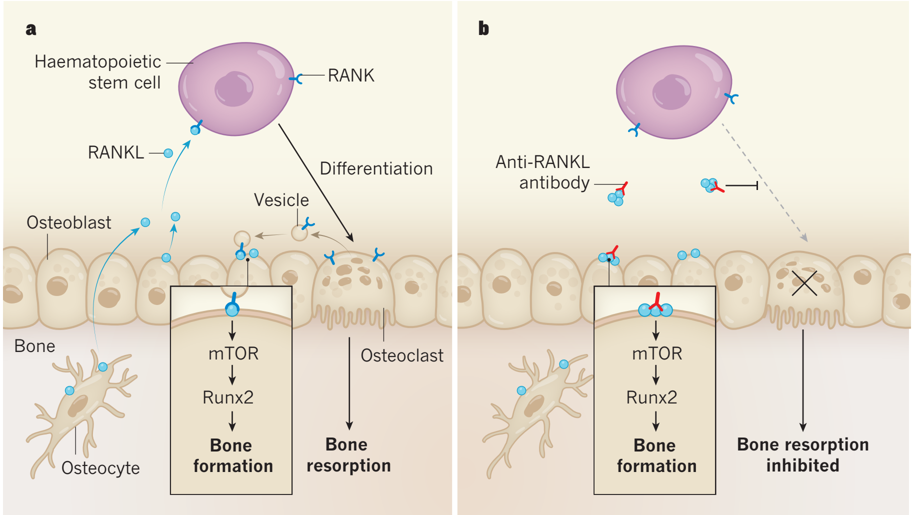


Figure 1 | Uncoupling bone formation and resorption. a, The signalling molecule RANKL is released by bone-producing cells called osteoblasts and their descendants, osteocytes. RANKL binds to the RANK receptor on haematopoietic stem cells, triggering their differentiation into cells called osteoclasts that resorb bone. Ikebuchi *et al.*[4] report that osteoclasts release extracellular vesicles that have RANK on their surface. Binding of vesicular RANK to RANKL on the surface of osteoblasts (and perhaps osteocytes, not shown) triggers intracellular signalling; the mTOR pathway is activated, which triggers production of proteins such as Runx2 that promote bone formation. **b,** The authors developed an anti-RANKL antibody, which binds and inactivates multiple RANKL monomers, thus reducing the formation of osteoclasts from their precursors. These antibodies also activate signalling in osteoblasts by binding cell-surface RANKL. Thus, they trigger bone formation while inhibiting bone resorption.

These intriguing data reveal a plausible physiological explanation for why RANK would be released from osteoclasts as cargo in extracellular vesicles[5]; they also establish that RANKL can act as a signal-transducing molecule at the cell surface, rather than merely as a secreted signalling protein. More broadly, this is the first demonstration of signal transduction by vesicular cell-surface receptors interacting with membrane-bound ligands. Finally, and perhaps most importantly, the study provides a mechanism by which osteoclasts in the process of resorbing bone communicate with nearby osteoblasts to regulate the extent to which new bone is formed. It remains unclear whether RANK-containing vesicles can activate cell-surface RANKL on osteocytes, which might not be in close proximity to osteoclasts.

On the therapeutic front, the anti-RANKL antibody denosumab is widely used to prevent and treat both osteoporosis and skeletal problems caused by the spread of cancers to bone. By preventing RANKL–RANK forward signalling, denosumab inhibits bone resorption by osteoclasts. But, in doing so, it transiently lowers bone formation, because of the tight coupling between osteoclasts and osteoblasts[7]. Ikebuchi and colleagues' newly engineered anti-RANKL antibody could potentially uncouple resorption and formation, and so be a more-effective alternative to denosumab. Notably, the authors found that, in mice whose ovaries had been removed (a model for post-menopausal osteoporosis), the antibody reduced bone resorption, but did not suppress bone formation.

It is difficult to speculate on whether Ikebuchi and co-workers' anti-RANKL antibody would have untoward side effects. This is certainly possible, because RANK and RANKL have many roles elsewhere in the body. For instance, immune cells called T cells express RANKL, and RANK is found on dendritic cells, with which T cells interact to trigger immune responses; in this setting, the interaction between RANK and RANKL enhances T-cell immunity[8]. The proteins have also been identified in the brain, notably in some neurons and neuron-supporting cells, in brain-specific immune cells called microglia, and in brain tissue deprived of oxygen[9]. They have also been implicated as mediators of fever-related responses to infections[10].

That said, a decade's worth of clinical studies using denosumab has revealed no discernible effects on the immune system or the brain. There is also no evidence yet for reverse RANK signalling in immune cells or neurons. Nonetheless, given that nearly all nucleus-bearing cells release extracellular vesicles, further studies of RANK–RANKL signalling in non-skeletal tissues are imperative. Given nature's propensity for reusing principles of intercellular signalling in many ways, the possibility of reverse signalling by other receptors must also be explored. ■

Mone Zaidi *and* **Christopher P. Cardozo** *are in the Mount Sinai Bone Program, Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, New York 10029, USA.*
C.P.C. *is also at the James J. Peters Veterans Affairs Medical Center, New York.*
e-mail: mone.zaidi@mountsinai.org

1. Zaidi, M. *Nature Med.* **13**, 791–801 (2007).
2. Nakashima, T. *et al. Nature Med.* **17**, 1231–1234 (2011).
3. Xiong, J. *et al. Nature Med.* **17**, 1235–1241 (2011).
4. Ikebuchi, Y. *et al. Nature* **561**, 195–200 (2018).
5. Huynh, N. *et al. J. Dent. Res.* **95**, 673–679 (2016).
6. Singha, U. K. *et al. J. Cell Biochem.* **103**, 434–446 (2008).
7. Zaidi, M., Yuen, T., Sun, L. & Rosen, C. J. *Endocr. Rev.* https://doi.org/10.1210/er.2018-00050 (2018).
8. Anderson, D. M. *et al. Nature* **390**, 175–179 (1997).
9. Shimamura, M. *et al. Proc. Natl Acad. Sci. USA* **111**, 8191–8196 (2014).
10. Hanada, R. *et al. Nature* **462**, 505–509 (2009).

This article was published online on 5 September 2018.

CONDENSED-MATTER PHYSICS

Spins travel far in an antiferromagnet

Controlled long-distance transport of electron spins is required for a kind of electronics known as spintronics. Such transport has been realized in an antiferromagnet, the most common type of magnetic material. SEE LETTER P.222

SERGIO M. REZENDE

Strongly magnetic materials called antiferromagnets have an intriguing property. Below a certain temperature, the spins (intrinsic angular momenta) of neighbouring atoms in the material point in opposite directions, so that the material exhibits no response to an external magnetic field. The discovery and pioneering studies of antiferromagnets were made by the physicist Louis Néel, who stated in his 1970 Nobel Lecture that the materials "are extremely interesting from the theoretical viewpoint, but do not seem to have any applications" (see go.nature.com/2lzrly8). This pessimistic view of antiferromagnets would change dramatically a few decades later. Today, the materials have practical applications, and promise to deliver several more. On page 222, Lebrun *et al.*[1] report the observation of long-distance spin transport in an antiferromagnetic insulator, demonstrating that such materials could be used for spin-based electronics (spintronics).

The debut of antiferromagnets in technology was made possible through the 1988 discovery of the giant magnetoresistance effect[2,3], which resulted in the 2007 Nobel Prize in Physics. This discovery showed that, in magnetic structures containing nanometre-thick multilayers (stacks of different ultrathin films), electron transport could be controlled by the spin of the electrons — rather than by their electric charge, as in conventional electronics. This finding triggered research into magnetic multilayers, and gave birth to the field of spintronics, which has revolutionized magnetic recording techniques and promises to bring about advances in information technologies[4].

Antiferromagnets proved to be essential in sensors that use the giant magnetoresistance effect. Since the late 1990s, the read heads for computer-disk drives have been based on such sensors. These read heads are much more sensitive to changes in magnetic fields than are conventional ones. However, although antiferromagnets are important for spintronic devices, they have had a passive role. Ferromagnets — materials in which all of the atomic spins are aligned — have had the active role.

In the past decade, this situation has begun to change because of experimental and theoretical results showing that antiferromagnets have several advantages over ferromagnets in spintronic devices. One advantage is the insensitivity of antiferromagnets to perturbations in external magnetic fields. Another is their ultrafast dynamics, which could enable devices to operate at terahertz-scale (10^{12} Hz) frequencies and therefore facilitate faster electronics. Developments in the past few years have given rise to the field of antiferromagnetic spintronics[5–7], which now gains a boost, thanks to Lebrun and colleagues.

The authors used a thin, flat sample of

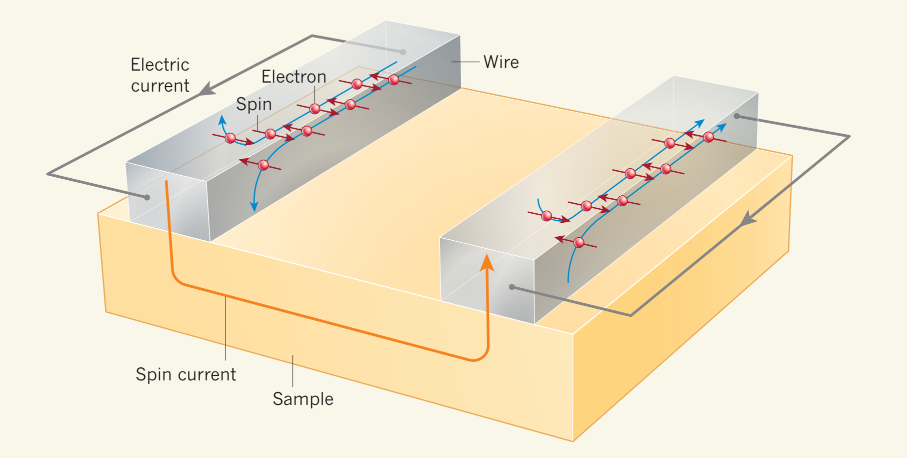


Figure 1 | Spin transport in an antiferromagnetic insulator. Lebrun *et al.*[1] carried out an experiment using a thin, flat sample of an electrically insulating material that exhibits a type of magnetism known as antiferromagnetism. The authors deposited two thin platinum wires on the surface of the sample and injected an electric current into one of the wires (the left wire). This electric current was converted into a spin current, which is produced by electrons with opposite spins (intrinsic angular momenta) moving in opposite directions (blue arrows). The spin current flowed into the sample and was transported laterally towards the second wire, where it was converted into an electric current. The authors demonstrated that the spin current could be transported over a relatively long distance between the wires (tens of micrometres at a temperature of 200 kelvin), which is a requirement for spin-based electronics.

single-crystal haematite, α-Fe_2O_3 — an electrical insulator that is the most common antiferromagnetic iron oxide. They deposited two thin platinum wires onto the surface of the sample, such that the two wires were parallel to each other, and injected an electric current into one of the wires (Fig. 1). This electric current was converted into a spin current, which in metals is produced by electrons with opposite spins moving in opposite directions, by means of a phenomenon called the spin Hall effect[8]. The spin current flowed into the sample and was transported laterally towards the second wire, where it was converted into an electric current by the inverse spin Hall effect[9]. Finally, this electric current produced a voltage signal that was picked up by a detector.

By reversing the direction of the injected electric current and calculating the difference in the voltages produced by currents in opposite directions, Lebrun *et al.* separated the signals arising from the spin Hall effect from those produced by thermal effects. This voltage difference varies linearly with the magnitude of the injected current and can be used to decode information carried by the spin current. The authors measured the voltage difference as a function of several parameters, such as the spatial separation of the wires, the intensity of an applied external magnetic field, and the angle between the direction of this field and the wires.

There are several key achievements in Lebrun and colleagues' work. First, the authors demonstrated spin transport in an antiferromagnetic insulator over a relatively long distance — tens of micrometres at a temperature of 200 kelvin. This finding suggests that such spin transport could be feasible at room temperature, which would be useful for practical applications. The authors argue that antiferromagnets have an advantage over the well-studied ferromagnetic insulator yttrium iron garnet ($Y_3Fe_5O_{12}$), which has been used in similar spin-transport experiments[10]. In devices based on this material, it is not possible to separate the signals produced by the spin Hall effect from those generated by thermal effects.

A second substantial achievement for device application is the authors' demonstration that the flow of the spin current can be controlled by an external magnetic field. The results also resolve a controversial issue: the physical mechanism responsible for spin transport in antiferromagnets. The dependence of the voltage difference on the spatial separation of the wires shows that the spin transport is mainly caused by the diffusion of excitations called magnons[11], rather than by the spins forming an exotic state of matter known as a superfluid[12].

Despite these achievements, some obstacles must be overcome before antiferromagnets can be used in practical devices requiring the control of spin currents. One challenge is pointed out by Lebrun and colleagues: finding antiferromagnets in which spins can be transported over even longer distances than demonstrated in the current work. Another, more formidable, challenge is to find materials amenable to spin-current control using weaker magnetic fields than those used by the authors. Nevertheless, the current findings provide additional impetus to the emerging field of antiferromagnetic spintronics. ■

Sergio M. Rezende *is in the Department of Physics, Federal University of Pernambuco, 50670-901 Recife, Pernambuco, Brazil.*
e-mail: rezende@df.ufpe.br

1. Lebrun, R. *et al. Nature* **561**, 222–225 (2018).
2. Baibich, M. N. *et al. Phys. Rev. Lett.* **61**, 2472–2475 (1988).
3. Binasch, G., Grünberg, P., Saurenbach, F. & Zinn, W. *Phys. Rev. B* **39**, 4828–4830 (1989).
4. Waldrop, M. M. *Nature* **530**, 144–147 (2016).
5. Jungwirth, T. *et al. Nature Phys.* **14**, 200–203 (2018).
6. Baltz, V. *et al. Rev. Mod. Phys.* **90**, 015005 (2018).
7. Jungfleisch, M. B., Zhang, W. & Hoffmann, A. *Phys. Lett. A* **382**, 865–871 (2018).
8. Hirsch, J. E. *Phys. Rev. Lett.* **83**, 1834–1837 (1999).
9. Saitoh, E., Ueda, M., Miyajima, H. & Tatara, G. *Appl. Phys. Lett.* **88**, 182509 (2006).
10. Cornelissen, L. J., Liu, J., Duine, R. A., Ben-Youssef, J. & van Wees, B. J. *Nature Phys.* **11**, 1022–1026 (2015).
11. Rezende, S. M., Rodríguez-Suárez, R. L. & Azevedo, A. *Phys. Rev. B* **93**, 054412 (2016).
12. Takei, S., Halperin, B. I., Yacoby, A. & Tserkovnyak, Y. *Phys. Rev. B* **90**, 094408 (2014).

BIOPHYSICS

A 3D cell shape found in sheets and tubes

Sheets of cells called epithelia can curve into tubes in embryos. Modelling and *in vivo* observations reveal that cells in tubes adopt an asymmetric cell shape dubbed scutoid, contrary to some previous assumptions.

GUY B. BLANCHARD

Embryo and organ development is driven in large part by the sculpting of sheets of cells called epithelia. Epithelial sheets have a remarkable repertoire of behaviours — they can narrow and elongate in plane, change thickness, and bend out of plane to form structures such as pits and furrows that develop into the body's tubes. However, our understanding of the 3D shapes of the cells that make up these structures, and of how they change during development, has been dominated by inferences often made from looking only at the most easily imaged surface of an epithelial-cell sheet. Writing in *Nature Communications*, Gómez-Gálvez *et al.*[1] use a modelling approach to predict how cells are packed to form tubes, and the authors give a name to a type of epithelial-cell shape involved in this packing, which they also find occurs *in vivo*.

Cells in single-layered epithelia have approximately parallel tops (apices) and bases. The cells are therefore usually thought of as prisms that join together in a honeycomb-like arrangement (Fig. 1a). When either the apices or the bases of cells constrict in all directions (isotropic constriction), wedge-shaped cells called frusta emerge (Fig. 1b), resulting in the formation of a pit[2] or furrow[3].

During the elongation of fruit-fly embryos, cell–cell junctions at the apices[4] or bases[5] of epithelial sheets can also contract in just one orientation in the plane of the sheet. This anisotropic constriction can lead to the formation of different contacts between neighbouring cells at the apical end of the cell from contacts at the basal end. Such constriction produces cells that are neither prisms nor frusta — but the shape that they form has had no name.

Gómez-Gálvez *et al.* analysed 3D cell shapes in epithelia by using computer modelling to study how cells pack into curved epithelia. For epithelia on a sphere, surface curvature is the same in all directions, and inverted frusta are expected. But the authors realized that, when epithelia form tubes and so curve in only one orientation, or form egg shapes that also have differential surface curvature, the cells' inner and outer surfaces can contact different neighbours. This change in contacts requires the formation of a triangular face on one side of the cell (Fig. 1c) — a 3D shape similar to that seen in the epithelial sheets of elongating fruit-fly embryos. The researchers name this 3D cell shape scutoid, from the Latin word for shield.

So why are scutoids predicted to occur in tubes? It is possible for 3D shapes resembling frusta to form tubes. But this requires the cells' apical or basal surfaces to be elongated, and Gómez-Gálvez *et al.* based their modelling on the plausible assumption that apical and basal surfaces of epithelial cells both tend towards isotropic shapes as the most energy-efficient ('relaxed') packing solution to differential curvature. This is in contrast to the sheets of fruit-fly embryos, in which the formation of scutoids is a dynamic, temporary process driven by active cell rearrangement[5]. When the authors modelled cells that had near-isotropic shapes on both the inner and outer surfaces

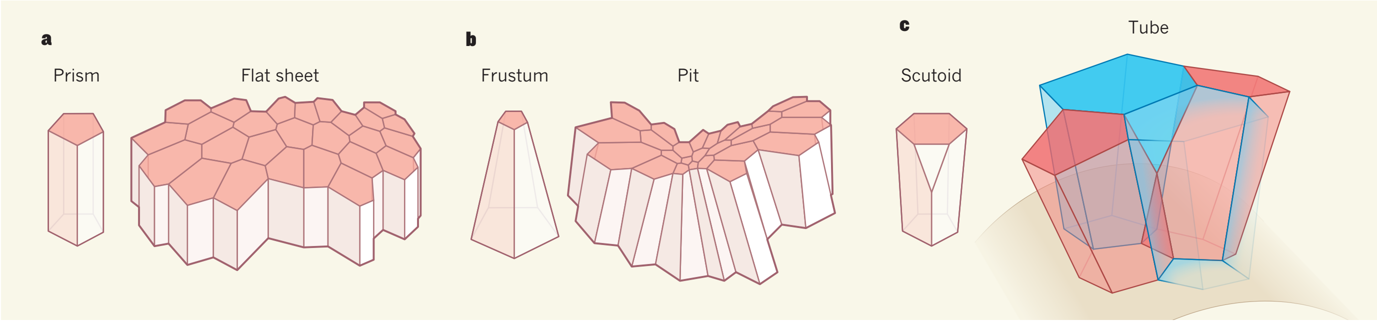


Figure 1 | 3D cell shapes in epithelial sheets. **a**, Sheets of cells called epithelia are most simply composed of prism shapes when flat. **b**, Frusta — prismatoids in which one end is constricted — bend epithelia into pits. **c**, Gómez-Gálvez *et al.*[1] provide evidence that cells can adopt a shape that the authors dub scutoid when epithelia form tubes. Scutoid cells have a triangular face towards one end and have different contacts with neighbouring cells at the inner and outer surfaces of the tube. For example, here the pink cells make contact only towards the inner surface of the tube. (Adapted from Fig. 1 of ref. 1.)

of tubes, they found topological differences between the inner and outer surfaces that required the cells to be scutoid. The authors then confirmed this finding using a mechanical model of the outer surface, in which the lengths of cell–cell interfaces were minimized, encouraging isotropic shapes.

Gómez-Gálvez and colleagues went on to search for scutoids *in vivo*. They found scutoid-shaped cells in snapshots of various embryonic epithelia, although, interestingly, not at the frequency predicted by their modelling. For example, the larval salivary glands of fruit flies have fewer scutoids than are predicted by the curvature of the gland tube. Furthermore, the authors found scutoids in the spherically curved surface layers of zebrafish embryos where none would be predicted.

There are several possible reasons for these discrepancies. First, scutoid predictions will be altered if they take into account other forces, either generated within the epithelium (for example, caused by anisotropic contractility) or acting on it. Second, in the current model, the authors used 2D modelling of the inner and outer surfaces of cells to infer 3D cell shapes — predictions might differ if modelling was extended to cell shapes that are explicitly 3D. However, a better understanding of the mechanics of the shared interfaces on the sides of cells will be needed before cell shapes more complex than prisms and frusta can be modelled accurately in three dimensions[6].

Third, the assumption that apical and basal cell shapes will tend towards isotropy is based on the assumption that the cells' cortex — a contractile, mesh-like network of proteins that gives cells their structure — acts as a ring at the cell–cell junctions at the ends of cells. But the cortex also spans the rest of the exposed apical and basal surfaces, and the contraction of this medial web seems not to tend to isotropic shapes in the same way[3]. Finally, scutoid formation might be highly dynamic in some epithelia, or might involve more-convoluted 3D shapes than those considered here, rendering measurement of scutoid frequency difficult.

Prisms, frusta and scutoids found in single-cell layered epithelia are discrete classes of shape, but belong in a continuum of cell shape and 3D arrangement. This continuum has recently been quantified as an additive combination of cell wedging and interleaving[2,7]. Wedging describes how cell shapes become wider or narrower with depth, whereas interleaving describes changes in how interdigitated a group of cells is with depth, equivalent to a rate of change of cell arrangement from bases to apices. Both wedging and interleaving can be quantified in units of radius of curvature, so are ideal for characterizing curved epithelia. A formal relationship between the degree of interleaving and the frequency of scutoids seems possible, which would make these continuous and discrete measures usefully interchangeable.

With Gómez-Gálvez and colleagues' characterization of the scutoid, we are beginning to recognize the kinds of 3D shapes and arrangements to look for in epithelia, and to develop the tools for quantifying them. Where else in nature should we expect scutoids? I would not bet against scutoids being found in plants, given the diversity of plant architectures. But because plant development is driven by cell growth and division without rearrangement, any mechanism of scutoid formation is likely to be rather different from that seen in animals. Nest comb construction by bees or wasps might also result in scutoids, particularly on curved surfaces, as in the nests of some paper wasps[8]. We shall have to wait and see. ■

Guy Blanchard *is in the Department of Physiology, Development and Neuroscience, University of Cambridge, Cambridge CB2 3DY, UK.*
e-mail: gb288@cam.ac.uk

1. Gómez-Gálvez, P. *et al. Nature Commun.* **9**, 2960 (2018).
2. Sanchez-Corrales, Y. E., Blanchard, G. B. & Röper, K. *eLife* **7**, e35717 (2018).
3. Chanet, S. *et al. Nature Commun.* **8**, 15014 (2017).
4. Rauzi, M., Lenne, P.-F. & Lecuit, T. *Nature* **468**, 1110–1114 (2010).
5. Sun, Z. *et al. Nature Cell Biol.* **19**, 375–383 (2017).
6. Bielmeier, C. *et al. Curr. Biol.* **26**, 563–574 (2016).
7. Deacon, W. J. PhD thesis, Univ. Cambridge (2012).
8. Wenzel, J. W. *J. Insect Behav.* **2**, 679–699 (1989).

ENVIRONMENTAL SCIENCE

The future of tidal wetlands is in our hands

Computational simulations suggest that future losses of tidal wetlands attributable to sea-level rise could be greatly offset by the landward advance of these ecosystems into newly sea-inundated areas. See Letter p.231

JONATHAN D. WOODRUFF

Coastal communities around the globe depend on tidal marshes and mangroves for the diverse ecological, economic and flood-mitigating services they provide. These relatively flat wetland systems (Fig. 1) commonly reside just above mean sea level, making them one of the ecosystems most at risk of being drowned by rising sea levels. But tidal wetlands will not disappear without a fight. On page 231, Schuerch *et al.*[1] present global-scale modelling that suggests that tidal wetlands are less vulnerable to sea-level rise than was thought. However, the scale of future wetland loss or gain depends greatly on the degree to which coastal communities accept or prevent the landward advances of these living coastal systems into newly inundated areas.

Tidal wetlands are dynamic, living systems

JOHANN SCHUMACHER/GETTY

Figure 1 | Tidal marshes at Jamaica Bay, New York. Schuerch and colleagues' modelling study[1] suggests that coastal management programmes could make a big difference to the future extent of tidal wetlands.

that have the ability to accelerate their growth as the sea level rises, using biophysical feedbacks — complex interactions between plant growth, water flow and sediment trapping. Furthermore, the flooding of low-lying areas by rising seas provides new locations that can accommodate tidal wetlands or allow them to grow. Are these adaptive mechanisms enough to save global wetlands from drowning in the future? Schuerch and colleagues' modelling study was framed to answer exactly this question.

The authors consider not only the adaptation potential of tidal wetlands to future sea-level rise in the vertical direction (upward growth) and horizontally (landward advance), but also the human response to their landward migration. Their modelling work was no small task, requiring the integration of large global data sets of shoreline topography, wetland distributions, concentrations of suspended sediments in wetland waters, tidal ranges (the vertical difference between high and low tides), and projected changes in sea level and coastal populations.

Schuerch *et al.* also incorporated a key feedback into their model to address the fact that the rate at which a wetland grows vertically to keep pace with sea-level rise is dictated not only by the plant matter it can produce, but also by how efficiently it traps sediment. So, in areas that have a sufficient sediment supply and tidal range, the model accounts for the acceleration of wetland growth that is due to the increased amount of sediment carried into and deposited on the wetland over a tidal cycle as sea levels rise.

As noted earlier, tidal wetlands can colonize new low-lying areas as they become flooded by sea-level rise, but little is known about the mechanisms that govern this inland migration. As a first cut, Schuerch and colleagues postulated that all newly inundated areas behind current wetlands and below a certain human population density will be converted to tidal wetlands — the second criterion is relevant because the size and extent of artificial barriers that impede inland wetland migration, such as roads, flood protection and other coastal infrastructure, are assumed to scale with population.

The authors tested what happens when the population threshold above which no wetland is created was set at either 5 or 20 people per square kilometre. These are considered to be the lower and upper bounds, respectively, of the population densities for which wetlands will migrate landward without any action being taken by humans to support or prevent the migration — that is, in 'business as usual' (BAU) scenarios. The authors also tested higher thresholds of 150 and 300 people per square kilometre, which they respectively describe as moderate and extreme nature-based adaptation scenarios, because they would require the replacement of current flood protection and infrastructure with alternatives that would allow wetlands to migrate into more-populated regions.

The authors' results suggest that global tidal wetland loss will be 0–30% by 2100 for the BAU scenarios in which wetlands migrate only into sparsely populated regions. These losses are well below those estimated in past studies[2–4], and highlight the need to consider adaptive feedbacks in future sustainability studies of tidal wetlands. Just as importantly, the nature-based adaptation scenarios suggest that wetland gains as high as 60% could be made when measures are taken that allow wetlands to migrate into more-populated areas.

Some of the findings are not too surprising. The study assumes an overly simple mechanism for inland wetland conversion, for which it is perhaps to be expected that the biggest increase in wetland extent will occur for the greatest sea-level rises under the most extreme nature-based adaptation scenarios. Other parts of the study are particularly interesting, however. For example, when the authors carried out simulations in which the concentration of sediments suspended in tidal wetland waters drops by 50% from present-day values, the extent of wetlands decreases by only 6%. Reductions in sediment supply caused by damming and other human activities have been a primary concern as contributors to wetland loss[5], but Schuerch and co-workers' results indicate that the degree to which wetlands are allowed to migrate into newly flooded lowland areas could have a much greater impact on wetland sustainability.

One notable caveat to the current findings is that the analysis assumes new wetlands are initiated at a fairly high tidal level. Such super-elevated wetlands will eventually drown if they lack a sufficient sediment supply to sustain them. However, the timescale for this drowning in the simulations was relatively long, and did not occur by 2100, the end of the modelled period.

Schuerch and colleagues' results highlight major gaps in our knowledge of wetland sustainability. Most of the modelled wetland gains were for mangrove systems, which currently represent about 70% of tidal wetlands[1]. However, our understanding of adaptive feedbacks in mangroves is poor compared with our understanding of tidal marshes. Data describing the inland migration of wetlands are also extremely limited, and environmental factors such as pre-existing soil and vegetative conditions could restrict migration to much lower extents than those projected by Schuerch and colleagues. However, the authors' study is a crucial step towards realistic assessments of future wetland changes, and highlights the key roles of both sea-level rise and nature-based adaptation strategies in providing new spaces where lowlands can sustainably accommodate the growth of tidal wetlands. ■

Jonathan D. Woodruff *is in the Department of Geosciences, University of Massachusetts, Amherst, Massachusetts 01003, USA.*
e-mail: woodruff@geo.umass.edu

1. Schuerch, M. *et al. Nature* **561**, 231–234 (2018).
2. Blankespoor, B., Dasgupta, S. & Laplante, B. *Ambio* **43**, 996–1005 (2014).
3. Spencer, T. *et al. Glob. Planet. Change* **139**, 15–30 (2016).
4. McFadden, L., Spencer, T. & Nicholls, R. J. *Hydrobiologia* **577**, 5–15 (2007).
5. Weston, N. B. *Estuaries Coasts* **37**, 1–23 (2014).

IMMUNOLOGY

Elusive inflammation insight uncovered

PINK1 and parkin proteins help to degrade damaged mitochondrial organelles, and abnormalities in these proteins are linked to Parkinson's disease. Mouse studies reveal that the proteins act to prevent inflammation. SEE LETTER P.258

ALEXANDRA STOLZ & IVAN DIKIC

Neurodegenerative conditions such as Parkinson's and Alzheimer's diseases constitute a major human-health burden. Although the symptoms, or the cells affected, can differ in such disorders, some neurodegenerative diseases have certain characteristics in common. These include a state of inflammation[1] and impaired elimination of defective mitochondrial organelles[2]. However, it remains to be determined whether such common alterations are interconnected, and whether they are a cause or a consequence of disease. On page 258, Sliter *et al.*[3] report their investigation of mice that have alterations in genes linked to Parkinson's disease. The authors identify a direct connection between the cellular process that eliminates damaged mitochondria — called mitophagy — and inflammation.

The enzymes PINK1 and parkin act in a pathway that attaches a protein called ubiquitin to cellular proteins; such ubiquitin-tagged components are targeted for cellular destruction. These enzymes assist with the process of mitophagy[4], in which non-functional mitochondrial fragments are rapidly sequestered into a membrane-bound vesicle that is degraded when it fuses with an organelle known as a lysosome.

Mutations that prevent the normal expression of PINK1 or parkin are linked to an early-onset form of Parkinson's disease[5], and there is evidence that failure to successfully Seeeliminate damaged mitochondria results in a higher risk of developing the disease[5]. However, mice that are deficient in PINK1 or parkin do not develop symptoms of the type observed in people who have abnormalities in the expression of these proteins; such symptoms include movement problems arising from the loss of neuronal cells that produce the neurotransmitter molecule dopamine[5,6]. Nor do these animals have the high level of inflammation that is a hallmark of Parkinson's disease[5,6].

The finding that the loss of PINK1 or parkin has a minimal effect on animals was surprising, because it was long thought that the removal of damaged mitochondria serves a key role in protecting cells from oxidative damage[5]. Defective mitochondria represent a severe threat to cells because ruptured mitochondria might release reactive oxygen species (ROS) that cause substantial cellular damage[6,7]. For example, ROS might increase the burden of potentially toxic

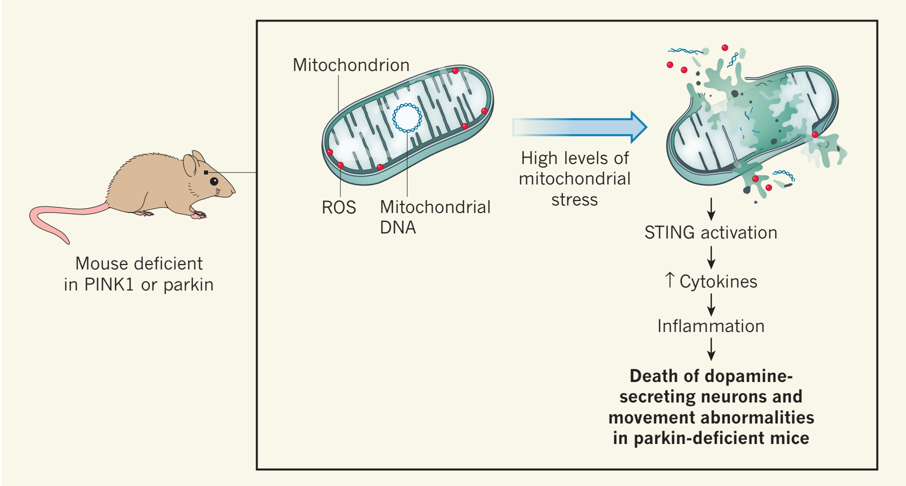

Mouse deficient in PINK1 or parkin
Mitochondrion
ROS
Mitochondrial DNA
High levels of mitochondrial stress
STING activation
↑ Cytokines
Inflammation
Death of dopamine-secreting neurons and movement abnormalities in parkin-deficient mice

Figure 1 | How the absence of PINK1 or parkin proteins leads to inflammation. Abnormalities in the proteins PINK1 or parkin are linked to early-onset Parkinson's disease in humans[5]. Mice that lack either protein are defective in the process that removes damaged organelles called mitochondria[5,6] in a controlled manner; this process is necessary to prevent organelle rupture and the release of reactive oxygen species (ROS) and mitochondrial DNA into the cytoplasm. However, these animals do not have the types of symptom found in human Parkinson's disease[5,6]. Sliter *et al.*[3] induced high levels of mitochondrial stress in these mice (by use of excessive levels of exercise or by a high level of mitochondrial-DNA mutations) and found that activation of the STING protein — which can mediate inflammation when mitochondrial DNA enters the cytoplasm — increases the expression of inflammation-inducing cytokine molecules. This indicates that PINK1 and parkin protect against inflammation, and might shed light on the inflammation that is commonly observed in people with Parkinson's disease[5,6]. In old mice that lack parkin, STING-mediated inflammation correlates with movement abnormalities and the loss of neuronal cells that secrete the neurotransmitter dopamine[10].

protein aggregates if proteins are subject to ROS-mediated damage. Defective mitochondria might also release components that are not normally present in the cytoplasm, such as mitochondrial DNA. Indeed, the intrusion of mitochondrial DNA into the cytoplasm can trigger inflammation[8,9] mediated by the protein STING. This raises the question of whether protection from inflammation, rather than from oxidative damage, might be the key role of mitophagy in the context of Parkinson's disease.

Sliter *et al.* investigated the consequences of PINK1 or parkin loss in mice that were subjected to a high level of mitochondrial stress. This stress was produced either by subjecting animals to an intensive, exhausting exercise regime or by exploiting a genetic alteration found in animals termed mutator mice — in which a defective polymerase enzyme causes a high level of mitochondrial-DNA mutations. It was previously reported[10] that old mutator mice that lack parkin have fewer dopamine-secreting neurons than normal, and that these mice develop movement abnormalities that are reminiscent of those observed in people who have Parkinson's disease. When the authors imposed mitochondrial stress on animals lacking PINK1 or parkin, they found that the bloodstream level of inflammation-driving molecules called cytokines was much higher than it was in mice that were not subjected to this mitochondrial stress.

However, the authors found that if mice lacked STING, as well as PINK1 or parkin, the expression of inflammatory cytokines did not increase as a result of mitochondrial stress. This indicated that STING is required to drive the inflammation mediated by this type of stress (Fig. 1). Moreover, an absence of STING prevented the movement defects and neuronal losses that usually occur in old mutator mice that lack parkin. The authors found that the bloodstream levels of the inflammatory cytokines IL-6, IL-1β and CCL2, which are elevated above normal in old mutator mice that lack parkin, are also higher than normal in people with Parkinson's disease who have mutations in both copies of the parkin gene. However, the authors observed that these cytokines were also elevated in disease-free relatives of people who have Parkinson's disease. Sliter and colleagues' study of these relatives, who have a mutation in only one of their two copies of the parkin gene, suggests that these particular cytokine alterations are not sufficient to cause the disease. Interestingly, people who receive long-term treatment with non-steroidal anti-inflammatory drugs have a lower than average risk of developing Parkinson's disease[11]. This observation is consistent with a model in which low levels of inflammation might protect against neurodegeneration.

The exciting results reported by Sliter and colleagues raise many important questions. How does STING-mediated inflammation cause neuronal death? Why are dopamine-secreting neuronal cells specifically affected? Is STING-dependent inflammation linked to other abnormalities associated with neurodegeneration, such as the formation of protein aggregates?

However, before these questions can be answered in the context of human disease, a crucial consideration is how well these mice provide a model of human Parkinson's disease. Further insights might come from using other systems, such as rats or fruit flies (*Drosophila*), which better mimic the types of change that occur in human Parkinson's disease. Finally, given that impaired mitophagy and inflammation are common features of several neurodegenerative disorders, it is tempting to speculate that STING-dependent inflammation might contribute in a similar way to other neurodegenerative conditions, such as Alzheimer's disease. It would be interesting to test this idea in experiments.

Sliter and colleagues' work might lead to studies in human cells that provide fresh insights into treating Parkinson's disease. Perhaps drugs that selectively inhibit STING-dependent inflammation will one day be used to treat or prevent disease — if it's possible to control any detrimental side effects on the immune system that might arise from targeting inflammation in this way. Will techniques such as monitoring the level of mitochondrial DNA in the bloodstream to detect abnormal mitophagy, or tracking the expression of STING-dependent cytokines, enable early diagnoses or make it possible to assess a person's risk of developing Parkinson's disease before symptoms appear? Sliter and colleagues' work points to new avenues of investigation in the efforts to improve the treatment options for Parkinson's disease. ■

Alexandra Stolz *and* **Ivan Dikic** *are in the Institute of Biochemistry II and Buchmann Institute for Molecular Life Sciences, Goethe University School of Medicine, 60590 Frankfurt am Main, Germany.*
e-mail: dikic@biochem2.uni-frankfurt.de

1. Glass, C. K., Saijo, K., Winner, B., Marchetto, M. C. & Gage, F. H. *Cell* **140**, 918–934 (2010).
2. Burté, F., Carelli, V., Chinnery, P. F. & Yu-Wai-Man, P. *Nature Rev. Neurol.* **11**, 11–24 (2015).
3. Sliter, D. A. *et al. Nature* **561**, 258–262 (2018).
4. Harper, J. W., Ordureau, A. & Heo, J. M. *Nature Rev. Mol. Cell Biol.* **19**, 93–108 (2018).
5. Pickrell, A. M. & Youle, R. J. *Neuron* **85**, 257–273 (2015).
6. Scarffe, L. A., Stevens, D. A., Dawson, V. L. & Dawson, T. M. *Trends Neurosci.* **37**, 315–324 (2014).
7. Yan, M. H, Wang, X. & Zhu, X. *Free Radic. Biol. Med.* **62**, 90–101 (2013).
8. Rongvaux, A. *et al. Cell* **159**, 1563–1577 (2014).
9. White, M. J. *et al. Cell* **159**, 1549–1562 (2014).
10. Pickrell, A. M. *et al. Neuron* **87**, 371–381 (2015).
11. Noyce, A. J. *et al. Ann. Neurol.* **72**, 893–901 (2012).

This article was published online on 22 August 2018.

MOLECULAR BIOLOGY

Proteins assemble as they are being made

An investigation finds that most protein complexes in yeast cells assemble before the subunits have fully formed. This mechanism might prevent the formation of toxic protein aggregates. SEE LETTER P.268

CHRISTINE MAYR

Most cellular processes are carried out by proteins, which generally assemble into heteromeric complexes — those composed of two or more distinct subunits. Although it was thought for many years that protein subunits diffuse freely in the cell and form complexes through random collisions, this seems unlikely, given that the cellular environment is extremely crowded. On page 268, Shiber *et al.*[1] provide *in vivo* evidence that, in eukaryotic organisms (which include animals, plants and fungi), most protein complexes in the cytoplasm are assembled co-translationally — that is, assembly occurs while at least one of the subunits is still being synthesized by the cell's ribosome machinery.

The study of co-translational protein-complex formation *in vivo* was challenging until a technique known as ribosome profiling was developed[2] in 2009. This technique allows the positions of ribosomes on messenger RNAs to be determined by sequencing RNA fragments, and is usually used to monitor translation — the process in which the ribosome decodes mRNA and uses it as a template for protein synthesis. Shiber *et al.* used a modified protocol called selective ribosome profiling[3], which isolates ribosomes that are synthesizing nascent protein chains already interacting with another protein. Subsequent sequencing of the corresponding RNA fragments reveals the mRNAs that encode the interacting nascent chains. The sequencing also identifies the protein domains involved in the interaction, because only ribosomes bound

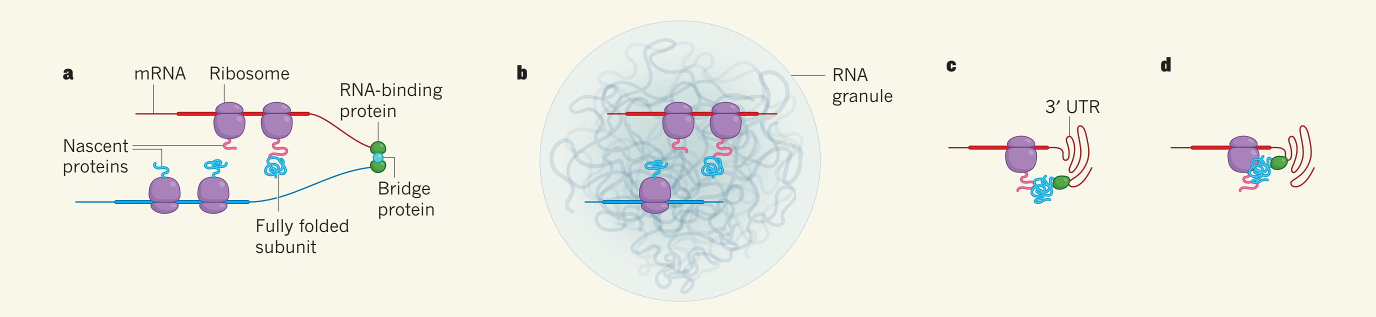


Figure 1 | Possible ways in which protein subunits are brought together for co-translational assembly. The cell's ribosome apparatus uses the sequences of messenger RNAs as templates for protein synthesis. Shiber *et al.*[1] report that the subunits of most cytoplasmic protein complexes in the yeast *Saccharomyces cerevisiae* are assembled co-translationally (that is, while at least one of the subunits is still being synthesized). But how are the different protein subunits brought together for assembly into complexes? **a,** One possibility is that an RNA-binding protein bridges two mRNAs, either directly (not shown) or indirectly (shown) through another protein. Here, a recently fully folded subunit formed on one mRNA has detached from its ribosome (which has disassembled; not shown), and interacts with a nascent protein on another mRNA. Thick regions of mRNAs represent translated regions; thin regions are not translated. **b,** The mRNAs could also come into close proximity in bodies known as RNA granules. **c,** An untranslated region of mRNAs known as the 3′ untranslated region (3′ UTR) might also recruit a fully folded subunit using an RNA-binding protein, bringing the subunit close to the mRNA's nascent protein. **d,** In a variant of **c**, the recruited subunit is temporarily deposited on the ribosome before being transferred to the nascent protein.

to nascent chains that contain fully exposed interaction domains will be isolated by selective ribosome profiling.

The authors investigated the prevalence of co-translational protein-complex assembly for 12 stable, well-characterized heteromeric complexes in the yeast *Saccharomyces cerevisiae.* They found that 9 of the 12 complexes assembled in a co-translational manner. Although it had been suspected that some protein complexes are assembled co-translationally[4,5], the finding that this process is widespread is surprising. Notably, the three protein complexes that did not seem to do this use dedicated chaperone proteins to assist in assembly. Given that the major function of chaperones is to prevent misfolding and random aggregation of proteins during protein folding[6], the researchers hypothesized that co-translational protein-complex assembly might serve a similar purpose.

Shiber *et al.* went on to find that six of the nine complexes that assemble co-translationally have a directional assembly mode: one of the subunits must be fully folded before it engages the nascent chain of a second subunit, but the fully folded second subunit cannot engage the nascent chain of the first subunit. This means that the second subunit must always participate in co-translational assembly as a nascent chain. Intriguingly, when the authors studied yeast strains that had been engineered not to produce the fully folded subunit, they observed that the nascent chain of the second subunit forms aggregates. This indicates that co-translational assembly does indeed prevent the formation of potentially toxic protein aggregates.

Although the authors convincingly show that co-translational protein-complex assembly is widespread, it is unclear how the subunits are brought into proximity to enable complex formation. There are two plausible, broad models, which are not mutually exclusive.

In the first model, the two mRNAs that encode the subunits are translated near to each other. This proximity could be achieved by an RNA-binding protein that bridges the mRNAs (Fig. 1a). Such a scenario has been suggested for the assembly of heteromeric ion channels[7]. However, a genome-wide analysis of human RNA–protein interactions[8] found that it is rare for two different mRNAs to be bridged by an RNA-binding protein. But a physical linkage is not necessarily needed to bring two mRNAs close together: mRNAs that share certain sequence motifs could co-localize in RNA bodies known as granules (Fig. 1b), thereby allowing them to be translated at defined subcellular locations[9].

In the second model, the two mRNAs are not close to each other. Instead, an RNA-binding protein recruits a fully folded protein subunit to an untranslated region (the 3′ UTR) of an mRNA that encodes the second subunit, thus allowing the folded subunit to interact with the second subunit as the latter is synthesized. In mammalian cells, 3′ UTRs of mRNAs have been shown to recruit proteins that then interact with the mRNAs' nascent or newly made proteins[10] (Fig. 1c). And in budding yeast, proteins recruited to mRNAs undergoing translation have been observed to be temporarily deposited on the ribosome before being transferred to the nascent protein chains (Fig. 1d)[11].

Recruitment of interacting proteins by 3′ UTRs is reminiscent of co-translational protein-complex formation in bacteria, in which protein subunits are often encoded by a cluster of neighbouring genes (an operon) that are expressed as a group[12,13]. However, in bacteria, the only fully folded protein that can interact with nascent proteins is one of the encoded subunits, whereas the use of the 3′ UTR in eukaryotes might allow several different interactors to be recruited to nascent proteins — thus enabling a variety of protein complexes to be assembled co-translationally[14].

More experiments are needed to work out how interacting protein subunits are brought together and how mis-assembly of complexes is prevented. Nevertheless, Shiber *et al.* have demonstrated that protein-complex formation often relies on recruitment mechanisms, rather than diffusion, to achieve specific protein interactions. Their findings add to an increasing number of *in vivo* observations suggesting that most cellular processes are interconnected: mRNAs not only encode proteins, but also increase the specificity of protein-complex formation by assisting the compartmentalization of proteins in the cytoplasm, and by regulating localized translation. Finally, it remains to be seen whether the majority of stable protein complexes in mammalian cells are also assembled co-translationally — but it seems likely that they are. ■

Christine Mayr *is in the Cancer Biology and Genetics Program, Memorial Sloan Kettering Cancer Center, New York, New York 10065, USA.*
e-mail: mayrc@mskcc.org

1. Shiber, A. *et al. Nature* **561**, 268–272 (2018).
2. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S. & Weissman, J. S. *Science* **324**, 218–223 (2009).
3. Becker, A. H., Oh, E., Weissman, J. S., Kramer, G. & Bukau, B. *Nature Protocols* **8**, 2212–2239 (2013).
4. Halbach, A. *et al. EMBO J.* **28**, 2959–2970 (2009).
5. Duncan, C. D. S. & Mata, J. *PLoS Genet.* **7**, e1002398 (2011).
6. Ellis, R. J. *Adv. Exp. Med. Biol.* **594**, 1–13 (2007).
7. Liu, F., Jones, D. K., de Lange, W. J. & Robertson, G. A. *Proc. Natl Acad. Sci. USA* **113**, 4859–4864 (2016).
8. Sugimoto, Y. *et al. Nature* **519**, 491–494 (2015).
9. Yasuda, K. *et al. J. Cell Biol.* **203**, 737–746 (2013).
10. Berkovits, B. D. & Mayr, C. *Nature* **522**, 363–367 (2015).
11. Chartron, J. W., Hunt, K. C. L. & Frydman, J. *Nature* **536**, 224–228 (2016).
12. Shieh, Y.-W. *et al. Science* **350**, 678–680 (2015).
13. Wells, J. N., Bergendahl, L. T. & Marsh, J. A. *Cell Rep.* **14**, 679–685 (2016).
14. Mayr, C. *Annu. Rev. Genet.* **51**, 171–194 (2017).

This article was published online on 29 August 2018.

BRIEF COMMUNICATIONS ARISING

Does the galaxy NGC1052–DF2 falsify Milgromian dynamics?

ARISING FROM P. van Dokkum et al. *Nature* **555**, 629–632 (2018); https://doi.org/10.1038/nature25767

A great challenge in current physics is to understand whether the observed internal dynamics of galaxies is due to dark matter or to a modification of the law of gravity. Recently, van Dokkum et al.[1] reported that the ultra-diffuse dwarf galaxy NGC1052–DF2 lacks dark matter, and they claimed that this would—paradoxically—be problematic for modified gravity theories such as Milgromian dynamics (MOND)[2,3]. However, NGC1052–DF2 is not isolated, so a valid prediction of its internal dynamics in MOND cannot be made without properly accounting for the external gravitational fields from neighbouring galaxies. Including this external field effect, following Haghi et al.[4], shows that NGC1052–DF2 is consistent with MOND.

In any viable cosmological model, both primordial and tidal dwarf galaxies, which form in gas-rich tidal debris when galaxies interact, should exist. Within the standard dark-matter-based cosmological model, primordial dwarfs are dark-matter-dominated, whereas tidal dwarf galaxies contain very little (if any) dark matter[5]. In MOND—a classical potential theory of gravity, derivable from a Lagrangian with conserved energy, total momentum and angular momentum[2,6]—the two types of dwarf galaxies cannot be distinguished. Until now, all known dwarf galaxies have shown similar dynamic behaviour, possibly implying falsification of the dark-matter models[7]. The discovery by van Dokkum et al.[1] of a galaxy lacking dark matter would thus constitute an important verification of the standard dark-matter cosmological model.

Given the density distribution of baryonic matter, ρ, the gravitational potential of Milgromian gravitation, ϕ, is determined by the generalized nonlinear Poisson equation[6] $\nabla\cdot[\mu(|\nabla\phi|/a_0)\nabla\phi] = 4\pi G\rho$, where the function $\mu(|\nabla\phi|/a_0)$ describes the transition from the Newtonian ($|\nabla\phi|/a_0 \gg 1$) to the Milgromian ($|\nabla\phi|/a_0 \ll 1$) regime[3,8], $a_0 = 3{,}600\ \mathrm{km^2\,s^{-2}\,kpc^{-1}}$ is Milgrom's constant[2,3] and G is Newton's universal constant of gravitation. An equation of this form has been used in investigations of classical models of quark dynamics and is therefore not without precedent in physics[6]. One implication of this equation is that the coupling strength of an object described by ρ depends on the position of another object, whereas its inertial mass is given only by its baryonic component. This breaking of the equivalence of the active gravitating and inertial mass of an object constitutes a prediction of a new physical phenomenon, which does not exist in standard gravitation models and may be visualized as the phantom dark-matter halo (incorporated in ϕ) being reduced in the presence of a sufficiently strong constant external field[9]. This may be related to the quantum vacuum[8]. A dwarf galaxy in the vicinity of a major galaxy may thus lose its phantom dark-matter halo and appear as a purely Newtonian system[3,10,11].

Following van Dokkum et al.[1], we assume that NGC1052–DF2 is located at a distance of $D = 20$ Mpc from Earth in the NGC1052 group, has a globular cluster population with an effective radius of $r_e = 3.1$ kpc (the stellar body has $r_e \approx 2.2$ kpc), an absolute V-band magnitude of $M_V = -15.4$ and a mass-to-light ratio of the stellar population of $M_{\rm DF2}/L_V = 2$, where $M_{\rm DF2}$ is the baryonic mass of NGC1052–DF2[1] and L_V is the absolute luminosity in the visual photometric band. If NGC1052–DF2 were isolated, then its MOND internal acceleration would be $a_i = 570\ \mathrm{km^2\,s^{-2}\,kpc^{-1}} = 0.15a_0$. This weak gravity would imply a MOND velocity dispersion of $\sigma_M \approx ((4/81)GM_{\rm DF2}a_0)^{1/4} \approx 20\ \mathrm{km\,s^{-1}}$. The 2σ (3σ) confidence range for the observed velocity dispersion of the globular clusters, calculated using the robust bi-weight estimator[12], is 0.0–14.2 km $\mathrm{s^{-1}}$ (0.0–19.7 km $\mathrm{s^{-1}}$), falsifying MOND with high significance[1].

This inference is only correct if NGC1052–DF2 is sufficiently isolated and at a distance of $D \approx 20$ Mpc. The bright elliptical host galaxy NGC1052 has a baryonic mass[13,14] of $M_{\rm NGC1052} \approx 10^{11} M_\odot$, where $M_\odot$ is the mass of the Sun. The projected separation between NGC1052–DF2 and NGC1052 is $D_{\rm sep} = 80$ kpc if both are at the same distance of $D = 20$ Mpc. At this separation, the acceleration induced by the MOND external-field of the host, $a_e = \sqrt{a_0 G M_{\rm NGC1052}}/D_{\rm sep}$ is $a_e \approx 0.14a_0$, such that $a_i \approx a_e \ll a_0$. Because the acceleration by the external field is as important as internal galactic acceleration, it needs to be properly taken into account in calculating the velocity dispersion in MOND, σ_M. This is true for any object of mass m, when its physical size is $r_e > (m/M)^{1/2} D_{\rm sep}$ and the external field is caused by another object of mass M at a distance of $D_{\rm sep}$.

Using a MOND N-body integrator[15], Haghi et al.[4] quantified the global one-dimensional line-of-sight velocity dispersion ($\sigma_{\rm M,EF}$) of a non-isolated stellar system lying in the intermediate external-field

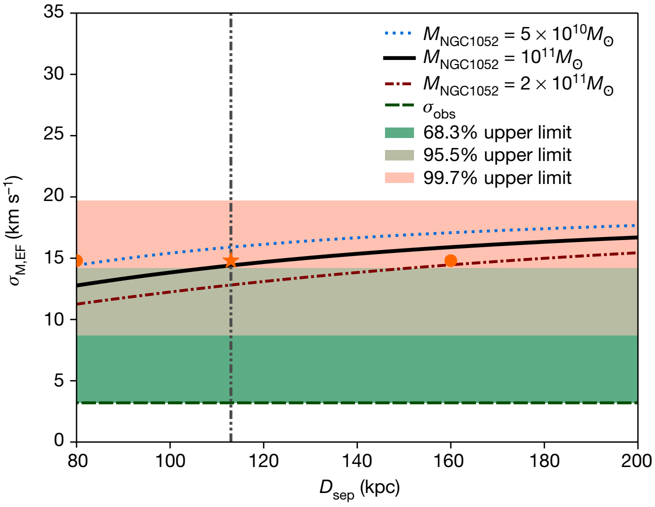


Fig. 1 | MOND velocity dispersion and external field. The line-of-sight MOND velocity dispersion, $\sigma_{\rm M,EF}$, of NGC1052–DF2 is shown as a function of the separation, $D_{\rm sep}$, between NGC1052–DF2 and NGC1052. The star (at $D_{\rm sep} = 113$ kpc, $M_{\rm NGC1052} \approx 10^{11} M_\odot$) and the two orange points (at $D_{\rm sep} = 80$ kpc, $M_{\rm NGC1052} \approx 5 \times 10^{11} M_\odot$ and at $D_{\rm sep} = 160$ kpc, $M_{\rm NGC1052} \approx 2 \times 10^{11} M_\odot$) correspond to the $a_e = 0.1a_0$ entry in table 1 of Haghi et al.[4] without interpolation. The vertical dash-dotted line indicates the more likely separation distance, $D_{\rm sep} = 2^{1/2} \times 80 = 113$ kpc. The 1σ, 2σ and 3σ upper confidence limits on the observed velocity dispersion[1], $\sigma_{\rm obs} = 3.2^{+5.5}_{-3.2}$ km $\mathrm{s^{-1}}$, are coloured as indicated in the key. The measured value is shown as the long-dashed horizontal line. Our interpolation of the values given in table 1 of Haghi et al.[4] (equation (1) in Supplementary Information) is shown for $M_{\rm NGC1052} = 5 \times 10^{10} M_\odot$, $10^{11} M_\odot$ and $2 \times 10^{11} M_\odot$ by the dotted, solid and dash-dotted lines, respectively.

regime (see Supplementary Information for more details). The dependence of the velocity dispersion of NGC1052–DF2, $\sigma_{\mathrm{M,EF}}$, on D_{sep} and M_{NGC1052} is shown in Fig. 1. The velocity dispersion declines from the isolated, asymptotic value $\sigma_{\mathrm{M}} = 20\ \mathrm{km\ s^{-1}}$ with decreasing D_{sep} and increasing mass because the external field of NGC1052 suppresses the phantom dark-matter halo, leading to Newtonian behaviour in the case of $a_{\mathrm{e}} \gg a_{\mathrm{i}}$. For the nominal host mass[14], $M_{\mathrm{NGC1052}} \approx 10^{11} M_{\odot}$, and for $D_{\mathrm{sep}} = 113$ kpc, the MOND velocity dispersion is in agreement within the 2σ confidence range of the observational value of van Dokkum et al.[1], as shown by the star in Fig. 1. Thus, the estimate of the internal velocity dispersion of the globular cluster system of NGC1052–DF2 according to MOND is $\sigma_{\mathrm{M,EF}} = 14\ \mathrm{km\ s^{-1}}$ for $M_{\mathrm{NGC1052}} \approx 10^{11} M_{\odot}$ and $D_{\mathrm{sep}} = 113$ kpc (Fig. 1), in agreement with the observed value within the 2σ confidence level. NGC1052–DF2 is thus consistent with an important MOND prediction.

For the external field to suppress the phantom dark-matter halo of NGC1052–DF2, this galaxy needs to be close to NGC1052. In van Dokkum et al.[1], the distance of NGC1052–DF2 is estimated using the surface-brightness fluctuation method. The calculations performed here show that NGC1052–DF2 falsifies MOND if its distance from Earth is $D > 18$ Mpc and its physical separation from NGC1052 or any other massive galaxy is $D_{\mathrm{sep}} > 300$ kpc. In this case, the external field becomes negligible and the galaxy has too small velocity dispersion for its stellar mass in MOND. On the other hand, the observed apparent luminosity of the galaxy is $l_{\mathrm{V}} = L_{\mathrm{V}} / D^2$. In Milgromian dynamics, $\sigma_{\mathrm{M}} \propto M_{\mathrm{DF2}}^{1/4} \propto L_{\mathrm{V}}^{1/4} \propto D^{1/2}$, so NGC1052–DF2 would be consistent with MOND at the 2σ confidence level if $D < 13$ Mpc, even if it were isolated. We note that the velocity dispersion predicted by Newtonian dynamics is also proportional to $D^{1/2}$ for a galaxy without dark matter. Future high-precision observations will ascertain the distance of NGC1052–DF2 and will constrain its gravitating mass using stellar velocities within NGC1052–DF2, instead of globular cluster velocities. As a word of caution, if this galaxy is not in virial equilibrium, then using the globular clusters as tracers for the mass may be unreliable, independently of the underlying cosmological model.

We acknowledge support from the Alexander von Humboldt Foundation, the German Academic Exchange Service (DAAD), the Swiss National Science Foundation, the Natural Science Foundation of China (grant numbers 11503025 and 11421303), the Anhui Natural Science Foundation (grant number 1708085MA20), the 'Hundred Talents Project of Anhui Province', the 'Fundamental Research Funds for the Central Universities', the 'DAAD-Ostpartnerschaftsprogramm für 2018' and the Iran Science Elites Federation (ISEF).

Data availability

All data are available from the corresponding author upon reasonable request.

Pavel Kroupa[1,2]*, Hosein Haghi[1,3], Behnam Javanmardi[4], Akram Hasani Zonoozi[1,3], Oliver Müller[5], Indranil Banik[6], Xufen Wu[7,8], Hongsheng Zhao[6,9] & Jörg Dabringhausen[2]

[1]Helmholtz-Institut für Strahlen- und Kernphysik, University of Bonn, Bonn, Germany. [2]Charles University in Prague, Faculty of Mathematics and Physics, Astronomical Institute, Prague, Czech Republic. [3]Institute for Advanced Studies in Basic Sciences, Department of Physics, Zanjan, Iran. [4]School of Astronomy, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran. [5]Departement Physik, Universität Basel, Basel, Switzerland. [6]Scottish Universities Physics Alliance, University of Saint Andrews, Saint Andrews, UK. [7]CAS Key Laboratory for Research in Galaxies and Cosmology, Department of Astronomy, University of Science and Technology of China, Hefei, China. [8]School of Astronomy and Space Science, University of Science and Technology of China, Hefei, China. [9]Department of Physics and Astronomy, LaserLaB, Vrije Universiteit, Amsterdam, The Netherlands. *e-mail: pkroupa@uni-bonn.de

Received: 4 April 2018; Accepted: 25 June 2018;
Published online 12 September 2018.

1. van Dokkum, P. et al. A galaxy lacking dark matter. *Nature* **555**, 629–632 (2018).
2. Milgrom, M. A modification of the Newtonian dynamics as a possible alternative to the hidden mass hypothesis. *Astrophys. J.* **270**, 365–370 (1983).
3. Famaey, B. & McGaugh, S. S. Modified Newtonian dynamics (MOND): observational phenomenology and relativistic extensions. *Living Rev. Relativ.* **15**, 10 (2012).
4. Haghi, H. et al. Testing fundamental physics with distant star clusters: theoretical models for pressure-supported stellar systems. *Mon. Not. R. Astron. Soc.* **395**, 1549–1557 (2009).
5. Ploeckinger, S. et al. Tidal dwarf galaxies in cosmological simulations. *Mon. Not. R. Astron. Soc.* **474**, 580–596 (2018).
6. Bekenstein, J. & Milgrom, M. Does the missing mass problem signal the breakdown of Newtonian gravity? *Astrophys. J.* **286**, 7–14 (1984).
7. Kroupa, P. The dark matter crisis: falsification of the current standard model of cosmology. *Publ. Astron. Soc. Aust.* **29**, 395–433 (2012).
8. Milgrom, M. The modified dynamics as a vacuum effect. *Phys. Lett. A* **253**, 273–279 (1999).
9. Milgrom, M. Solutions for the modified Newtonian dynamics field equation. *Astrophys. J.* **302**, 617–625 (1986).
10. McGaugh, S. S. MOND prediction for the velocity dispersion of the "feeble giant" Crater II. *Astrophys. J.* **832**, L8 (2016).
11. Thomas, G. F. et al. Stellar streams as gravitational experiments. II. Asymmetric tails of globular cluster streams. *Astron. Astrophys.* **609**, A44 (2018).
12. Beers, T. C., Flynn, K. & Gebhardt, K. Measures of location and scale for velocities in clusters of galaxies - a robust approach. *Astron. J.* **100**, 32–46 (1990).
13. Fernández-Ontiveros, J. A., López-Sanjuan, C., Montes, M., Prieto, M. A. & Acosta-Pulido, J. A. The most recent burst of star formation in the massive elliptical galaxy NGC 1052. *Mon. Not. R. Astron. Soc.* **411**, L21–L25 (2011).
14. Bellstedt, S. et al. The SLUGGS survey: a comparison of total-mass profiles of early type galaxies from observations and cosmological simulations, to 4 effective radii. *Mon. Not. R. Astron. Soc.* **476**, 4543–4564 (2018).
15. Londrillo, P. & Nipoti, C. N-MODY: a code for collisionless *N*-body simulations in modified Newtonian dynamics. *Mem. Soc. Astron. Ital. Suppl.* **13**, 89 (2009).

Author contributions P.K. conceived the work and wrote the manuscript; H.H. and A.H.Z. performed the external field effect calculations; O.M. and B.J. contributed observational expertise on galaxies; I.B., X.W. and H.Z. contributed MOND expertise; J.D. contributed dwarf galaxy expertise. All co-authors contributed to the manuscript with comments and passages.

Competing interests Declared none.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0429-z.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to P.K.

https://doi.org/10.1038/s41586-018-0429-z

https://doi.org/10.1038/s41586-018-0483-6

Optimized arylomycins are a new class of Gram-negative antibiotics

Peter A. Smith[1,15]*, Michael F. T. Koehler[2,15], Hany S. Girgis[1], Donghong Yan[3], Yongsheng Chen[4], Yuan Chen[5], James J. Crawford[2], Matthew R. Durk[5], Robert I. Higuchi[6,11], Jing Kang[3], Jeremy Murray[7], Prasuna Paraselli[6,12], Summer Park[3], Wilson Phung[8], John G. Quinn[9], Tucker C. Roberts[6,13], Lionel Rougé[7], Jacob B. Schwarz[2,14], Elizabeth Skippington[10], John Wai[4], Min Xu[3], Zhiyong Yu[4], Hua Zhang[3], Man-Wah Tan[1] & Christopher E. Heise[1,9]*

Multidrug-resistant bacteria are spreading at alarming rates, and despite extensive efforts no new class of antibiotic with activity against Gram-negative bacteria has been approved in over fifty years. Natural products and their derivatives have a key role in combating Gram-negative pathogens. Here we report chemical optimization of the arylomycins—a class of natural products with weak activity and limited spectrum—to obtain G0775, a molecule with potent, broad-spectrum activity against Gram-negative bacteria. G0775 inhibits the essential bacterial type I signal peptidase, a new antibiotic target, through an unprecedented molecular mechanism. It circumvents existing antibiotic resistance mechanisms and retains activity against contemporary multidrug-resistant Gram-negative clinical isolates in vitro and in several in vivo infection models. These findings demonstrate that optimized arylomycin analogues such as G0775 could translate into new therapies to address the growing threat of multidrug-resistant Gram-negative infections.

The discovery and development of several classes of safe and efficacious antibiotics has markedly reduced mortality from bacterial infections. However, the overuse and misuse of these same antibiotics—both in medicine and in agriculture—has driven the rapid evolution and dissemination of antibiotic resistance[1]. The 'ESKAPE' pathogens (*Enterococcus faecium*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa* and *Enterobacter* species)[2] present the most acute threat of developing untreatable multidrug-resistant (MDR) infections, with the Gram-negative members of this group (*Escherichia coli*, *K. pneumoniae*, *P. aeruginosa* and *A. baumannii*) posing a particular threat owing to their dual-membrane envelope that prevents many antibiotics from accessing their targets. Despite considerable effort, no new class of antibiotic has been approved for Gram-negative pathogens in over fifty years.

The arylomycins are a class of macrocyclic lipopeptides[3,4] that inhibit the bacterial type I signal peptidase (SPase), an essential membrane-bound protease that uses an atypical serine–lysine dyad to cleave signal sequences from pre-proteins following their translocation across the cytoplasmic membrane[5]. SPase has been pursued as an antibiotic target for nearly twenty years[6–9], with a focus on developing agents with activity against Gram-positive bacteria, in which the active site of the enzyme is exposed on the surface of the cell. In Gram-negative bacteria the active site of SPase is located in the periplasmic space between the cytoplasmic membrane and the outer membrane, and was thought to be inaccessible to the arylomycins because their high molecular weight and lipophilicity would preclude penetration of the outer membrane[10–12]. Consistent with this hypothesis, no member of the arylomycin class has shown activity against the Gram-negative ESKAPE pathogens. It has been demonstrated that arylomycins have some intrinsic ability to penetrate the outer membrane, and that their lack of activity against Gram-negative pathogens is attributable in part to a naturally occurring mutation in the Gram-negative SPase LepB that reduces the binding affinity of arylomycin[13,14]. Encouraged by these results, we initiated a systematic optimization effort to identify arylomycin analogues with increased target affinity and improved penetration of the outer membrane. Here we describe the discovery of G0775, a synthetic arylomycin derivative with potent in vitro antibacterial activity against the Gram-negative ESKAPE pathogens. Notably, the physicochemical properties of G0775 remain outside the range that is currently considered conducive to yielding potent Gram-negative activity, which suggests that it uses an atypical mechanism to penetrate the outer membrane. Crucially, highly MDR pathogens that are resistant to nearly all available antibiotic therapies are susceptible to G0775, and de novo resistance to G0775 occurs at a low frequency. The potent in vitro activity of G0775 translates into robust in vivo efficacy in several infection models, demonstrating the potential of these optimized natural products to address the growing clinical threat of antibiotic resistant Gram-negative bacteria.

Optimized arylomycins have Gram-negative activity

Arylomycins consist of a macrocyclic tripeptide core with an N-terminal lipopeptide tail (purple), a C-terminal carboxylic acid (green) and two phenols on the aromatic rings of the macrocycle (red), as shown in Fig. 1. With the aim of modifying this natural scaffold into a broadly active Gram-negative antibiotic, we explored modifications at all three regions. Because crystal structures of the Gram-negative SPase LepB in complex with arylomycin indicate that the macrocyclic core preorganizes the peptide backbone and occupies a conserved region of the substrate-binding pocket[15,16], we chose to leave this region unchanged. Naturally occurring variants of the arylomycin N-terminal lipopeptide tail are known to modulate the spectrum of activity against

---

[1]Department of Infectious Diseases, Genentech, South San Francisco, CA, USA. [2]Department of Discovery Chemistry, Genentech, South San Francisco, CA, USA. [3]Department of Translational Immunology, Genentech, South San Francisco, CA, USA. [4]Department of Chemistry, Wuxi AppTec, Shanghai, China. [5]Department of Drug Metabolism and Pharmacokinetics, Genentech, South San Francisco, CA, USA. [6]RQx Pharmaceuticals, La Jolla, CA, USA. [7]Department of Structural Biology, Genentech, South San Francisco, CA, USA. [8]Department of Microchem Proteomics, Genentech, South San Francisco, CA, USA. [9]Department of Biochemical and Cellular Pharmacology, Genentech, South San Francisco, CA, USA. [10]Department of Bioinformatics and Computational Biology, Genentech, South San Francisco, CA, USA. [11]Present address: Retrovirox, San Diego, CA, USA. [12]Present address: Vertex, San Diego, CA, USA. [13]Present address: Aduro Biotech, Berkeley, CA, USA. [14]Present address: FLX Bio, South San Francisco, CA, USA. [15]These authors contributed equally: Peter A. Smith, Michael F. T. Koehler. *e-mail: smith.peter@gene.com; heise.christopher@gene.com

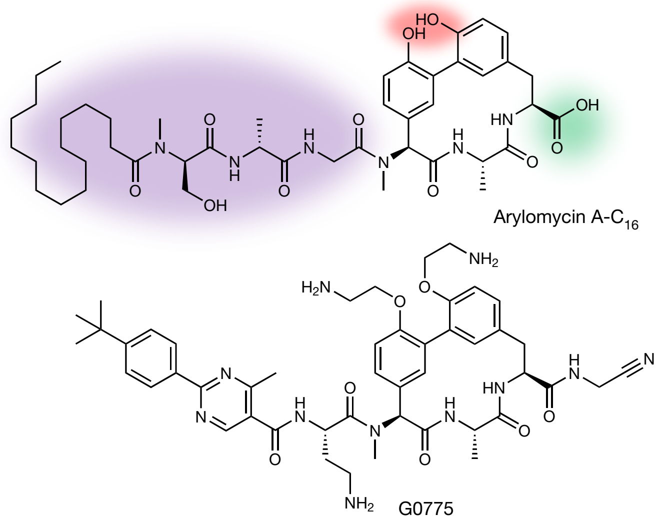


Fig. 1 | Chemical structures of arylomycin A-C$_{16}$ and G0775. Purple indicates the N-terminal lipopeptide tail, red indicates the location of phenolic oxygen modifications to the arylomycin core macrocycle and green is the site for appendage of the nitrile electrophile that covalently binds to the LepB catalytic lysine.

Table 1 | MIC values for arylomycin A-C$_{16}$ and G0775 against pathogenic Gram-negative bacteria

Bacterial species	MIC (μg ml^{-1})	
	Arylomycin A-C$_{16}$	G0775
E. coli ATCC 25922	>64	0.125
K. pneumoniae ATCC 43816	>64	0.125
Enterobacter cloacae ATCC 13407	>64	0.5
Enterobacter aerogenes ATCC 13408	>64	0.125
Citrobacter werkmanii ATCC 51114	>64	0.25
Serratia marcescens CDC4385-74	>64	0.5
A. baumannii ATCC 17978	>64	1
P. aeruginosa ATCC 27853	>64	2

Gram-positive bacteria[5,13,17,18]. The N-terminal lipopeptide tail also interacts with the amino acid residue in LepB that reduces the binding affinity of the arylomycin natural products[13], so we first searched for variants that more optimally engage this region of the Gram-negative protein. Ultimately we achieved the greatest improvement in activity by shortening the linear D-*N*-Me-Ser-D-Ala-Gly tripeptide in arylomycin to a single diaminobutyric acid and replacing the natural aliphatic tail with 2-(4-(*tert*-butyl)phenyl)-4-methylpyrimidine-5-carboxylic acid. This modification considerably improved the Gram-negative activity, yielding a molecule with measurable minimum inhibitory concentrations (MICs) against *E. coli* and *K. pneumoniae* (Extended Data Fig. 6). Second, we appended electrophilic warheads at the C-terminal carboxylate in an attempt to covalently bind the catalytic serine of LepB[14]. We identified 2-aminoacetonitrile as the optimal modification, which further improved activity against *E. coli* and *K. pneumoniae* and also imparted measurable activity against the particularly challenging bacteria *A. baumannii* and *P. aeruginosa* (Extended Data Fig. 6). Finally, inspired by modifications to the macrocyclic phenols found in natural arylomycin analogues and previous reports that phenol modifications can afford activity gains[14,19], we explored the simultaneous substitution of both phenols (Fig. 1). The resulting molecule, termed G0775, is at least 500-fold more potent than arylomycin A-C$_{16}$ against the ESKAPE pathogens *E. coli* and *K. pneumoniae* as well as against related pathogens from the same family. Moreover, the potent activity of G0775 extends to the notoriously difficult-to-treat non-fermenting Gram-negative bacteria *P. aeruginosa* and *A. baumannii* (Table 1). Although not the primary focus of our efforts, G0775 also exhibits markedly increased potency against the important Gram-positive pathogens methicillin-resistant *S. aureus* and *Staphylococcus epidermidis*, with MICs for each less than 0.5 μg ml^{-1} (Extended Data Fig. 6). G0775 therefore validates our hypothesis that chemical modifications can be used to expand the spectrum and potency of the natural product arylomycins to include ESKAPE pathogens and represents a new and exciting molecule that could address the need for new Gram-negative antibiotics.

G0775 is active against MDR bacteria

After demonstrating potent activity against common laboratory strains of Gram-negative ESKAPE pathogens, we set out to determine whether G0775 circumvents the multitude of resistance mechanisms that undermine the effectiveness of currently available therapies. First, we determined the MIC of G0775 against a panel of 49 MDR clinical isolates of *E. coli* and *K. pneumonia*, which are provided by the Centers for Disease Control and Prevention (CDC) to evaluate potential new antibiotics. The panel is highly enriched for MDR pathogens, and more than half exhibited resistance to at least five of the eight clinically used antibiotics that were tested (Supplementary Table 1). In stark contrast to the approved drugs, G0775 maintains its potent activity against all 49 isolates, with 90% of the isolates exhibiting MICs of $\leq$0.25 μg ml^{-1} (Supplementary Table 1). G0775 also remains active against 16 highly MDR strains of *A. baumannii* and 12 highly MDR strains of *P. aeruginosa* from the CDC, with 90% of the isolates exhibiting MICs of $\leq$4 μg ml^{-1} and $\leq$16 μg ml^{-1}, respectively. To better assess the ability of G0775 to circumvent specific resistance mechanisms, we focused on a single MDR isolate of *K. pneumoniae*, strain CDC 0106. Whole-genome sequencing of CDC 0106 revealed at least 10 chromosomally encoded and 25 plasmid-encoded genes associated with resistance to 13 classes of antibiotics (Supplementary Table 2). These genetic elements encoded by CDC 0106 confer resistance to nearly all available antibiotic classes (Table 2). Moreover, the presence of genes encoding the NDM-1 metallo beta-lactamase and the ArmA ribosomal methylase is expected to confer resistance even to new therapies recently approved for clinical use such as Vabomere and Zemdri, respectively[20,21], which are new variants of known antibiotic classes. G0775 remains highly active against CDC 0106, confirming that this new class is not affected by this extensive repertoire of resistance elements.

G0775 acts via LepB with low resistance frequency

We next set out to confirm that the improved antibacterial activity of G0775 was achieved through specific inhibition of LepB and not by means of off-target or secondary mechanisms. We first created a strain in which LepB expression levels are controlled by the concentration of arabinose in the growth medium (Extended Data Fig. 1), and found that low levels of LepB expression resulted in a 16-fold increase in susceptibility relative to wild-type cells, whereas high levels of LepB expression decrease susceptibility by eightfold (Extended Data Table 1). Next, we explored the rate at which G0775 kills bacteria. Genetic depletion of LepB in *E. coli* leads to time-dependent cell death[22], and G0775 likewise causes time-dependent killing at a similar rate, ultimately reducing the number of viable bacteria by more than six orders of magnitude after 24 h (Extended Data Fig. 1). The target titration and time-kill data are consistent with G0775 acting via LepB inhibition. To unequivocally confirm this, we assessed the ability of diverse Gram-negative ESKAPE pathogens to spontaneously evolve resistance to G0775. As expected for an inhibitor that targets a single gene product[23], resistant mutants were isolated at a moderate frequency in the presence of a low concentration of G0775 (4× MIC). When selections were performed with a higher concentration of G0775 (8× MIC), the frequency of resistance was below the limit of detection for *K. pneumoniae* and *A. baumannii*. At 16× MIC, the frequency of resistance for *E. coli* and *P. aeruginosa* decreased to less than 10^{-10} (Fig. 2a). Targeted and whole-genome sequencing of G0775-resistant *E. coli* mutants revealed that resistance-conferring mutations in all but four isolates were found

Table 2 | MIC values for G0775 and antibiotics from diverse classes against MDR *K. pneumoniae* CDC 0106

Class	Antibiotic	MIC (μg ml^{-1})
Cephalosporin/β-lactamase inhibitor	Ceftazidime/avibactam	>64
Aminoglycoside	Gentamicin	>64
Fluoroquinolone	Levofloxacin	>64
Carbapenem	Meropenem	>64
Polymyxin	Colistin	16
Tetracycline	Tigecycline	2
Arylomycin	G0775	0.5

within the substrate-binding groove of LepB, which confirms that the whole-cell activity of G0775 is on target. Encouragingly, these mutations reduce susceptibility to G0775 by only 4- to 16-fold (Extended Data Table 2), which is consistent with the marked decrease in resistance frequencies in the presence of higher drug concentrations. The remaining four isolates each contained a missense mutation (N282Y) in *acrB*, a gene that encodes an efflux pump; this may increase the affinity of this efflux system for G0775. Deletion of the AcrB or TolC efflux pumps in wild-type *E. coli* had no effect on the potency of G0775, which suggests that AcrB(N282Y) is a gain-of-function mutation and demonstrates that the native *E. coli* efflux systems do not considerably affect the activity of G0775 (Extended Data Table 1). Taken together, these results demonstrate that G0775 kills Gram-negative bacteria by means of LepB inhibition, and that spontaneous resistance to G0775 occurs primarily via on-target mutations that provide only modest decreases in susceptibility.

G0775 binds to LepB in an unprecedented manner

To test the hypothesis that resistance-conferring *lepB* mutations alter residues that directly contact the inhibitor, we solved the co-crystal structure of LepB in complex with G0775 to 2.4 Å resolution. By mapping the selected G0775-resistant mutations onto the *E. coli* LepB crystal structure, we confirmed that these mutations lie within the substrate-binding groove and closely overlay the G0775 binding site (Fig. 2b). Indeed, all of the resistance-conferring mutations lie within 4.5 Å of the bound inhibitor and, in total, mutations were identified in 8 of the 20 residues that contact G0775. Overall the LepB–G0775 structure reveals that the macrocyclic core and the backbone amide heteroatoms of the inhibitor bind in an almost identical manner to that previously observed in structures of arylomycins bound to LepB[15,16], despite the addition of an electrophilic warhead to the C terminus (Extended Data Fig. 2). Notably, the aminoacetonitrile warhead on G0775 does not engage the catalytic serine (S91) as we had expected, but instead reacts with the nitrogen of the catalytic lysine (K146) (Fig. 3a, Extended Data Fig. 4). This is consistent with the formation of a covalent amidine through direct nucleophilic attack of the electrophilic nitrile by the catalytic lysine, which is an inversion of the typical roles of the active-site dyad (Extended Data Fig. 3) and, to our knowledge, is the first report of amidine formation between an inhibitor and its target protein. To confirm the nature of this covalent interaction, we incubated LepB with an excess of G0775 overnight, digested with trypsin and subjected the digest to analysis by liquid chromatography coupled with mass spectrometry (LC–MS) (Fig. 3b). The resulting fragments confirm the formation of an adduct of the expected molecular weight between G0775 and a LDYIKR peptide fragment that includes the catalytic residue K146. This mechanism of covalent protease inhibition is—to our knowledge—unique to our molecules, and adds a new tool to the rapidly expanding field of covalent enzyme inhibitors[24]. Detailed kinetic enzymology studies of full-length LepB embedded in a detergent micelle were performed to quantify the rate of covalent-bond formation with the catalytic lysine. The data indicate very tight reversible binding of the molecule into the substrate-binding pocket ($K_I = 0.44$ nM) followed by irreversible inactivation ($k_{\text{inact}} = 0.0007$ s^{-1}; $k_{\text{inact}}/K_I = 1{,}590{,}909$ M^{-1} s^{-1}) (Fig. 3c), which confirms the very high affinity of G0775 for LepB.

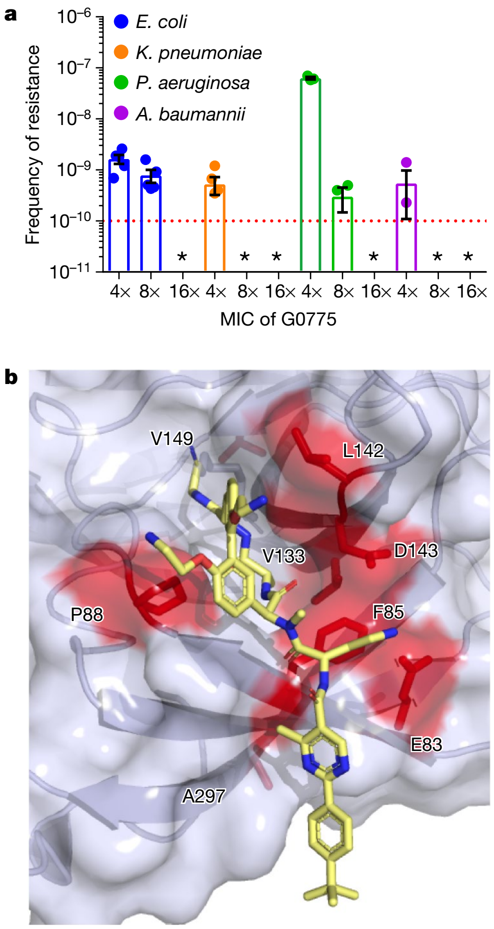


Fig. 2 | Resistance to G0775. a, Frequencies of resistance of *E. coli* ATCC 25922, *K. pneumoniae* 43816, *P. aeruginosa* PAO1 and *A. baumannii* ATCC 17978 to G0775 at 4, 8 and 16 times the respective MIC of each strain. The limit of detection (10^{-10}) is marked by the red dotted line, with asterisks representing data points that fall below the limit. Data shown for *E. coli* and *K. pneumoniae* are an average of five independent experiments, whereas data shown for *P. aeruginosa* and *A. baumannii* are an average of three independent experiments. Error bars represent s.d. **b**, LepB target mutations mapped onto the G0775–LepB crystal structure. Mutants were spontaneously generated in *E. coli* ATCC 25922 by overnight plating on G0775 at 4× MIC and are indicated in red. G0775 is shown in stick representation.

G0775 access to LepB is porin-independent

To inhibit LepB, which is localized on the inner membrane, G0775 must first cross the outer membrane. As a surrogate for the efficiency of outer membrane penetration, we determined the potency of G0775 against a hyperpermeable *E. coli* mutant, *lptD4213* (*imp*), which exhibits a defect in the outer membrane biogenesis pathway and is highly sensitive to both lipophilic and large hydrophilic antibiotics that are otherwise excluded by the outer membrane[25]. The *imp* mutation renders *E. coli* 32-fold more sensitive to G0775, and EDTA-mediated chelation (EDTA, ethylenediaminetetraacetic acid) of the divalent cations that are necessary to maintain the stability of the outer membrane results in a similar increase in potency (Extended Data Table 1). These data suggest that, despite its high molecular mass and high polarity, G0775 penetrates the outer membrane with reasonable efficiency. The majority of polar antibiotics that cross the outer membrane do so through open, water-filled channels called porins. However, deletion of OmpC or OmpF, the primary porins in *E. coli*, either individually or in combination has no effect on the potency of G0775 (Extended Data Table 1). Although it remains possible that other porins contribute to outer membrane penetration, porin mutations were never identified in resistance studies and the high molecular mass of G0775 (890 g mol^{-1})

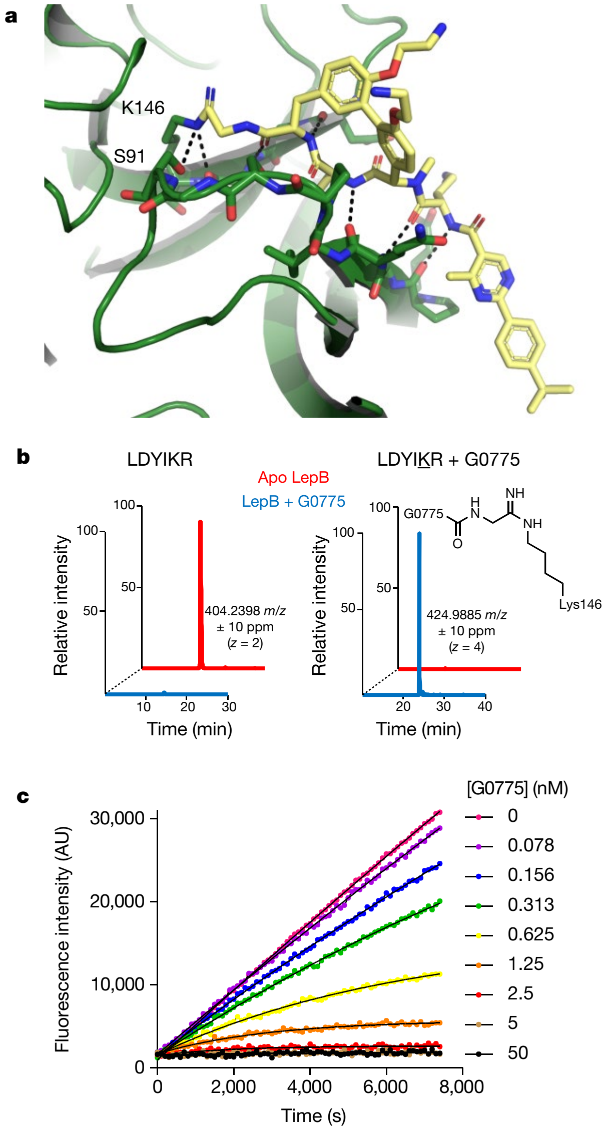


Fig. 3 | G0775 binds the LepB protease domain to form an irreversible covalent bond with catalytic lysine 146. a, Crystal structure of the protease domain of LepB at 2.4 Å resolution, with G0775 represented as sticks covalently bound to lysine 146. **b**, LC–MS detection of LDYIKR LepB peptide fragment after tryptic digest following the incubation of LepB with G0775. The unmodified peptide (red) is detected only in the absence of incubation with G0775, whereas the LDYIKR–G0775 peptide adduct (blue) is detected only after the incubation of LepB with G0775. This study was performed twice with similar results. **c**, LepB kinetic enzyme assays in the presence of the indicated concentrations of G0775. The nonlinear inactivation curves indicate time-dependent inhibition consistent with covalent-bond formation. K_{inact} (0.0007 ± 0.0002 s^{-1}) and K_{I} (0.44 ± 0.15 $\text{M}^{-1}\ \text{s}^{-1}$) were measured from three independent experiments and the data points shown are averages of four replicates from a single experiment. AU, arbitrary units.

is above the range typically considered compatible with porin-mediated entry[26] (Extended Data Fig. 5c). An alternative entry mechanism, whereby compounds bind to and destabilize the outer membrane, is termed 'self-promoted uptake', and the presence of positively charged groups can facilitate this process[27]. An analysis of the activity of more than 1,000 distinct arylomycin analogues against wild-type and outer-membrane-permeable *E. coli* suggests a modest beneficial effect of amines on the outer membrane penetrance of arylomycins (Extended Data Fig. 5a, b). To assess more directly the contribution of positive charge to the outer membrane penetrance of G0775, we synthesized neutral analogues that lacked all three primary amines (Extended Data Figs. 5b, 6). These analogues remain active against wild-type *E. coli*, which demonstrates that amines are not essential for outer membrane penetrance; however, consistent with our global analysis, the efficiency of these analogues in penetrating the outer membrane, as measured in our shift assay, is eightfold lower than that of G0775. Notably, removal of the amines from G0775 reduces potency against both outer-membrane-permeable *E. coli* and against the Gram-positive bacteria *S. aureus* by 32- to 64-fold, which is consistent with an important role for the amines in the intrinsic potency of G0775. Taken together, these data suggest that G0775 and related analogues penetrate the outer membrane through a porin-independent mechanism, which is likely to be a form of self-promoted uptake that is enhanced by the presence of positive charge.

Efficacy of G0775 in vivo against ESKAPE pathogens

The potent in vitro activity of G0775 in the presence of 50% serum or 1% lung surfactant (Extended Data Table 1) suggested that this antibiotic could be used to treat serious systemic and pulmonary infections. To confirm this, we tested G0775 in several pre-clinical infection models with different pathogenic Gram-negative bacteria. First, we evaluated the in vivo efficacy of G0775 in the murine neutropenic thigh infection model, which assesses intrinsic antibacterial activity independent from the innate immune system and has established translatability to efficacy in humans. G0775 demonstrates potent bactericidal activity (>2-log decrease in colony-forming units) against the *E. coli* strain ATCC 25922 and the *K. pneumoniae* strain ATCC 43816 at 1 mg kg^{-1} and 5 mg kg^{-1}, respectively (Fig. 4a, Extended Data Fig. 7). Next, we tested G0775 against *P. aeruginosa* ATCC 27853 and *A. baumannii* ATCC 17978 in the same neutropenic thigh model. A higher concentration of G0775 was required for efficacy, consistent with the higher MIC values for these pathogens (Fig. 4a). Third, we used a lung-infection model to assess the activity of G0775 against an MDR bacterial pulmonary infection. Using the MDR CDC 0106 (Table 2) as the challenge pathogen, G0775 was bacteriostatic at 2 mg kg^{-1} and bactericidal at 20 mg kg^{-1} (Fig. 4b). These results confirm that this new class of antibiotic is able to overcome all of the resistance mechanisms inherent to CDC 0106 in vitro and in vivo. Finally, we examined the ability of G0775 to protect mice from a lethal challenge of *K. pneumoniae* Z strain ATCC 43816 in a mucin peritonitis model. G0775 effected a dose-dependent increase in survival, with 5 mg kg^{-1} delivered twice on day 0 of the infection able to maintain full viability of mice until 84 h (Fig. 4c). Taken together, these in vivo efficacy data, combined with a lack of mammalian cell toxicity (Extended Data Table 3), demonstrate the potential of G0775 to treat several types of infections caused by a variety of wild-type and MDR Gram-negative bacteria.

Discussion

The growth in MDR bacterial infections is a serious risk to global health, and most antibacterial discovery efforts focus on reinvigorating existing classes of clinically approved antibiotics by means of chemical modification. Although this approach has been critical and effective, the resulting molecules are eventually subject to the same types of resistance mechanisms, which evolve and disseminate in response to previously used members of the class. By contrast, owing to its novel mechanism of action, G0775 bypasses existing resistance mechanisms and exhibits potent in vitro and in vivo activity against MDR Gram-negative bacteria, including extremely drug-resistant strains such as the clinical isolate CDC 0106.

The potent whole-cell activity of G0775 is the product of excellent affinity for its target, the avoidance of efflux and its relatively efficient outer membrane penetration. Ultimately, the ability of G0775 to cross the outer membrane is surprising, given that its physicochemical properties are quite distinct from those of other approved Gram-negative antibiotics[10]. This disconnect probably reflects the fact that most Gram-negative antibiotics cross the outer membrane through porins, whereas G0775 uses a porin-independent mechanism. Although considerable advances have been made in our understanding of porin-mediated

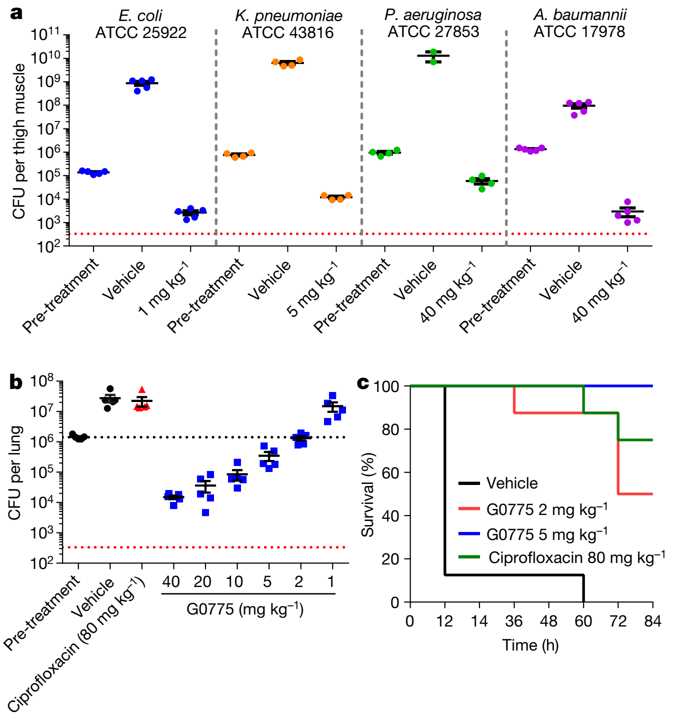


Fig. 4 | In vivo efficacy of G0775. a, Thigh infections initiated in neutropenic mice with the indicated Gram-negative bacterial species were treated with G0775 ($n = 5$ mice for *E. coli* and *A. baumannii*; $n = 4$ mice for *K. pneumoniae* and *P. aeruginosa*) or vehicle ($n = 2$ mice for *P. aeruginosa*, $n = 4$ mice for *K. pneumoniae*, $n = 5$ mice for *E. coli* and *A. baumannii*), and the bacterial burden was quantified 20 h after infection. G0775 was delivered subcutaneously twice during the infection period at 2 and 11 h post-infection at the indicated dose. **b**, Dose-dependent antibacterial activity of G0775 against a bacterial infection established by the MDR *K. pneumoniae* clinical isolate CDC 0106 in the lungs of neutropenic mice ($n = 5$). G0775 was delivered subcutaneously twice during the 20-h infection period at 2 and 11 h after the initiation of infection. The centre bars for **a** and **b** represent the mean, error bars represent s.e.m. and dotted red lines represent the limit of determination of bacterial colony forming units (CFU). **c**, Kaplan–Meier survival analysis of mice after peritoneal infection with *K. pneumoniae* ATCC 43816. G0775 was delivered subcutaneously twice on day zero at 2 and 11 h after the initiation of infection, and the viability of mice ($n = 8$) was monitored until 84 h.

entry[11,12]—culminating in prospective guidelines for improving the porin-mediated accumulation of compounds in Gram-negative bacteria[28]—considerably less is known about the features that enable efficient porin-independent penetration of the outer membrane[29]. Polymyxins, the polycationic antibiotics of last resort, cross the outer membrane by means of self-promoted uptake owing to their positive charge[27], and resistance readily emerges through alterations to the outer membrane that block this process[30]. Importantly, G0775 remains active against polymyxin-resistant isolates, and our data show that positive charge is beneficial to, but not essential for, the penetration of arylomycin through the outer membrane. Indeed other large, polar, non-charged inhibitors of periplasmic targets—such as globomycin—have Gram-negative antibacterial activity and exhibit relatively efficient penetration of the outer membrane[31]. A better understanding of the features that enable efficient porin-independent penetration of the outer membrane may help to inform the design of new Gram-negative antibiotics with properties that are distinct from current classes, and our ongoing studies of G0775 and related molecules may help to elucidate these features.

In summary, the broad spectrum and potent activity of G0775, combined with its low vulnerability to spontaneous resistance and excellent preclinical efficacy, suggest that optimized arylomycin analogues may represent a much-needed new class of Gram-negative antibiotic. Such a new mechanistic class of antibiotic has the potential to reset the clock in the ongoing arms race with pathogenic bacteria and help to postpone the prospect of a return to the pre-antibiotic era.

Reporting summary

Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The sequence reported in this paper has been deposited in the GenBank database, www.ncbi.nlm.nih.gov/genbank (accession numbers CP022611–CP022613). G0775, G3031, G8126 and G6850 are unique molecules proprietary to Genentech that are readily available on request. The authors declare that all other data supporting the findings of this study are available from the authors and/or included with the manuscript as Source Data (efficacy studies) or as Supplementary Information.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0483-6.

Received: 6 October 2017; Accepted: 26 June 2018;
Published online 12 September 2018.

1. Ventola, C. L. The antibiotic resistance crisis part 1: causes and threats. *P&T* **40**, 277–283 (2015).
2. Rice, L. B. Federal funding for the study of antimicrobial resistance in nosocomial pathogens: no ESKAPE. *J. Infect. Dis*. **197**, 1079–1081 (2008).
3. Schimana, J. et al. Arylomycins A and B, new biaryl-bridged lipopeptide antibiotics produced by *Streptomyces* sp. Tü 6075. I. Taxonomy, fermentation, isolation and biological activities. *J. Antibiot. (Tokyo)* **55**, 565–570 (2002).
4. Kulanthaivel, P. et al. Novel lipoglycopeptides as inhibitors of bacterial signal peptidase I. *J. Biol. Chem*. **279**, 36250–36258 (2004).
5. Smith, P. A., Powers, M. E., Roberts, T. C. & Romesberg, F. E. *In vitro* activities of arylomycin natural-product antibiotics against *Staphylococcus epidermidis* and other coagulase-negative staphylococci. *Antimicrob. Agents Chemother*. **55**, 1130–1134 (2011).
6. Paetzel, M., Dalbey, R. E. & Strynadka, N. C. The structure and mechanism of bacterial type I signal peptidases. A novel antibiotic target. *Pharmacol. Ther*. **87**, 27–49 (2000).
7. Smitha, R. C. & Anné, J. Bacterial type I signal peptidases as antibiotic targets. *Future Microbiol*. **6**, 1279–1296 (2011).
8. Bruton, G. et al. Lipopeptide substrates for SpsB, the *Staphylococcus aureus* type I signal peptidase: design, conformation and conversion to α-ketoamide inhibitors. *Eur. J. Med. Chem*. **38**, 351–356 (2003).
9. Payne, D. J., Gwynn, M. N., Holmes, D. J. & Pompliano, D. L. Drugs for bad bugs: confronting the challenges of antibacterial discovery. *Nat. Rev. Drug Discov*. **6**, 29–40 (2007).
10. O'Shea, R. & Moser, H. E. Physicochemical properties of antibacterial compounds: implications for drug discovery. *J. Med. Chem*. **51**, 2871–2878 (2008).
11. Nikaido, H. Molecular basis of bacterial outer membrane permeability revisited. *Microbiol. Mol. Biol. Rev*. **67**, 593–656 (2003).
12. Brown, D. G., May-Dracka, T. L., Gagnon, M. M. & Tommasi, R. Trends and exceptions of physical properties on antibacterial activity for Gram-positive and Gram-negative pathogens. *J. Med. Chem*. **57**, 10144–10161 (2014).
13. Smith, P. A., Roberts, T. C. & Romesberg, F. E. Broad-spectrum antibiotic activity of the arylomycin natural products is masked by natural target mutations. *Chem. Biol*. **17**, 1223–1231 (2010).
14. Liu, J., Smith, P. A., Steed, D. B. & Romesberg, F. Efforts toward broadening the spectrum of arylomycin antibiotic activity. *Bioorg. Med. Chem. Lett*. **23**, 5654–5659 (2013).
15. Luo, C., Roussel, P., Dreier, J., Page, M. G. P. & Paetzel, M. Crystallographic analysis of bacterial signal peptidase in ternary complex with arylomycin A2 and a β-sultam inhibitor. *Biochemistry* **48**, 8976–8984 (2009).
16. Liu, J. et al. Synthesis and characterization of the arylomycin lipoglycopeptide antibiotics and the crystallographic analysis of their complex with signal peptidase. *J. Am. Chem. Soc*. **133**, 17869–17877 (2011).
17. Roberts, T. C., Schallenberger, M. A., Liu, J., Smith, P. A. & Romesberg, F. E. Initial efforts toward the optimization of arylomycins for antibiotic activity. *J. Med. Chem*. **54**, 4954–4963 (2011).
18. Therien, A. G. et al. Broadening the spectrum of β-lactam antibiotics through inhibition of signal peptidase type I. *Antimicrob. Agents Chemother*. **56**, 4662–4670 (2012).
19. Peters, D. S., Romesberg, F. E. & Baran, P. S. Scalable access to arylomycins via C–H functionalization logic. *J. Am. Chem. Soc*. **140**, 2072–2075 (2018).
20. Castanheira, M., Huband, M. D., Mendes, R. E. & Flamm, R. K. Meropenem-vaborbactam tested against contemporary gram-negative isolates collected worldwide during 2014, including carbapenem-resistant, KPC-producing, multidrug-resistant, and extensively drug-resistant *Enterobacteriaceae*. *Antimicrob. Agents Chemother*. https://doi.org/10.1128/AAC.00567-17 (2017).
21. Endimiani, A. et al. ACHN-490, a neoglycoside with potent *in vitro* activity against multidrug-resistant *Klebsiella pneumoniae* isolates. *Antimicrob. Agents Chemother*. **53**, 4504–4507 (2009).

22. Lüke, I., Handford, J. I., Palmer, T. & Sargent, F. Proteolytic processing of *Escherichia coli* twin-arginine signal peptides by LepB. *Arch. Microbiol*. **191**, 919–925 (2009).
23. Silver, L. L. Multi-targeting by monotherapeutic antibacterials. *Nat. Rev. Drug Discov*. **6**, 41–55 (2007).
24. Singh, J., Petter, R. C., Baillie, T. A. & Whitty, A. The resurgence of covalent drugs. *Nat. Rev. Drug Discov*. **10**, 307–317 (2011).
25. Sampson, B. A., Misra, R. & Benson, S. A. Identification and characterization of a new gene of *Escherichia coli* K-12 involved in outer membrane permeability. *Genetics* **122**, 491–501 (1989).
26. Silver, L. L. Challenges of antibacterial discovery. *Clin. Microbiol. Rev*. **24**, 71–109 (2011).
27. Hancock, R. E. & Bell, A. Antibiotic uptake into Gram-negative bacteria. *Eur. J. Clin. Microbiol. Infect. Dis*. **7**, 713–720 (1988).
28. Richter, M. F. et al. Predictive compound accumulation rules yield a broad-spectrum antibiotic. *Nature* **545**, 299–304 (2017).
29. Richter, M. F. & Hergenrother, P. J. The challenge of converting Gram-positive-only compounds into broad-spectrum antibiotics. *Ann. NY Acad. Sci*. **24**, 71 (2018).
30. Raetz, C. R. H., Reynolds, C. M., Trent, M. S. & Bishop, R. E. Lipid A modification systems in Gram-negative bacteria. *Annu. Rev. Biochem*. **76**, 295–329 (2007).
31. Kiho, T. et al. Structure–activity relationships of globomycin analogues as antibiotics. *Bioorg. Med. Chem*. **12**, 337–361 (2004).

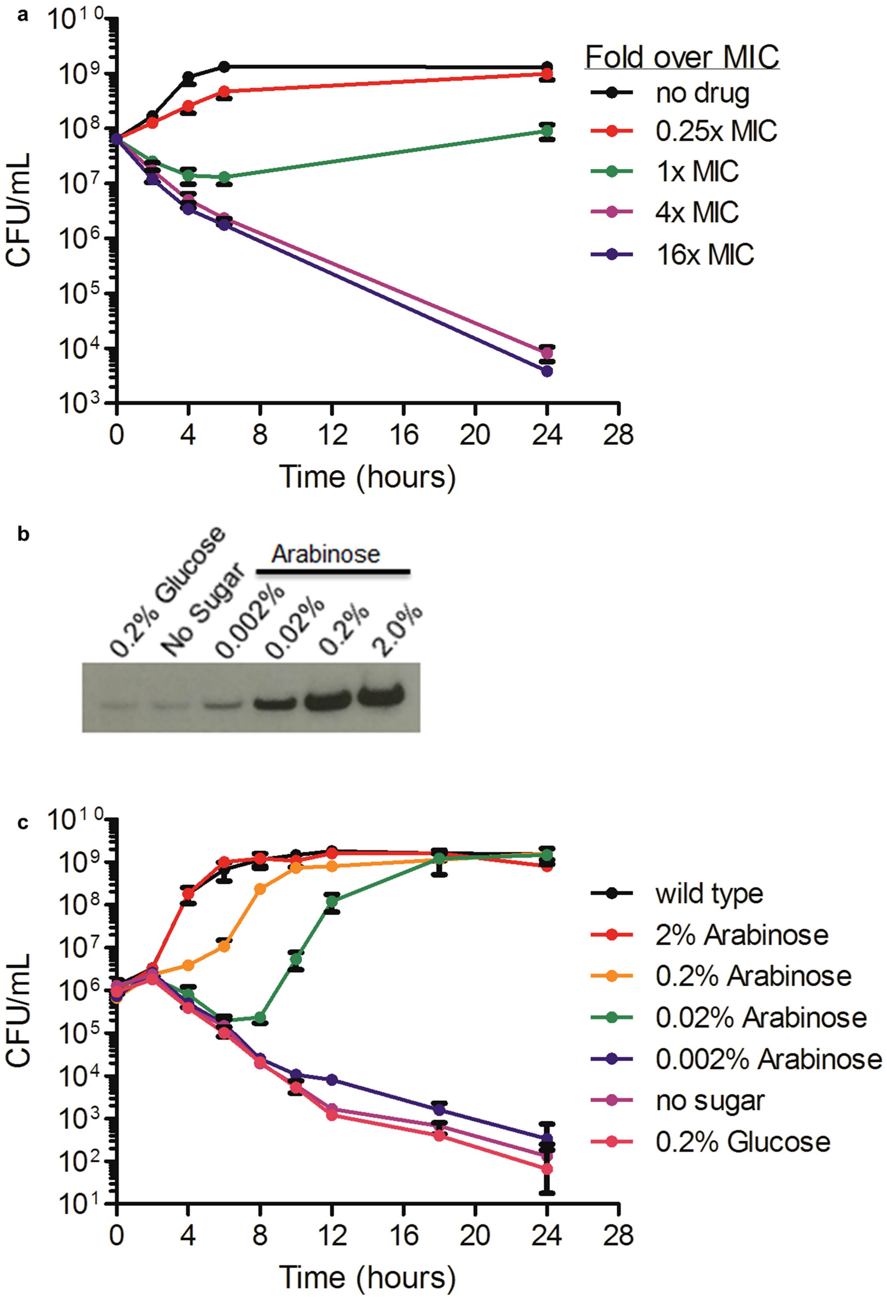
Acknowledgements We thank Confluence Discovery Technologies for running the enzymatic inactivation studies and WuXi Shanghai for determining MIC values to support the chemistry effort. We thank C. Peng, H. La, J. Z. Chen, F. Ma, X. Liang, S. Ubhayakar and the protein expression group from Genentech for supporting the research efforts required to discover G0775, J. Guillory for performing the whole-genome sequencing on CDC 0106, and B. Sellers for performing property calculations on synthesized compounds.

Reviewer information *Nature* thanks E. Brown, P. Hergenrother, T. Maier and the other anonymous reviewer(s) for their contribution to the peer review of this work.

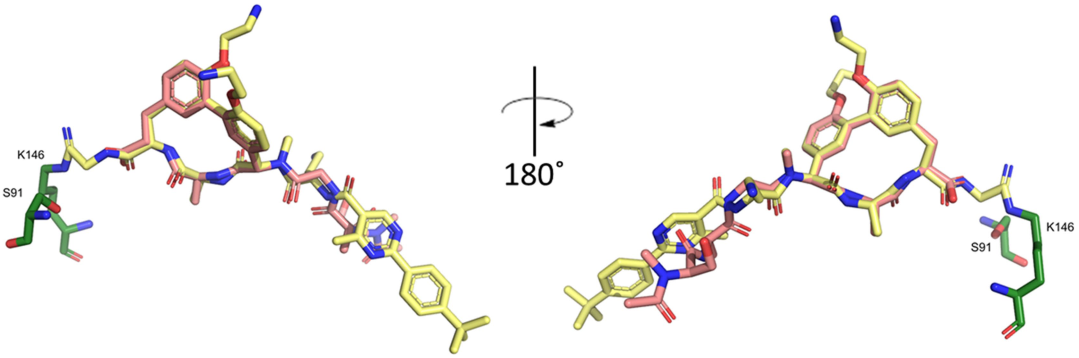
Author contributions P.A.S., M.F.T.K., Yu.C., M.R.D., M.X., M.W.T. and C.E.H. conceived the study and managed the execution of experiments. H.S.G. and P.A.S. performed all the microbiology experiments. D.Y., J.K., S.P., M.X. and H.Z. performed all in vivo efficacy studies. J.M. and L.R. performed all crystallographic experiments. W.P. performed the proteomics experiments. C.E.H. and J.G.Q. performed enzymology experiments. M.F.T.K., P.A.S., Yo.C., J.J.C., T.C.R., R.I.H., P.P., J.B.S., J.W. and Z.Y. designed and synthesized the molecules. E.S. performed the whole-genome sequencing. P.A.S., M.F.T.K. and C.E.H. wrote the manuscript with input from all authors.

Competing interests P.A.S., M.F.T.K., Yo.C., J.J.C., R.I.H., P.P., T.C.R., J.B.S. and Z.Y are listed as inventors on the pending patent application WO2017084630, which covers the molecules G0775, G6850, G3031 and G8126. P.A.S., M.F.T.K., Yu.C., M.R.D., M.X., M.W.T., H.S.G., D.Y., J.K., S.P., H.Z., J.M., L.R., W.P., J.G.Q., J.J.C. and C.E.H. declare active employment by Genentech at the time of submission.

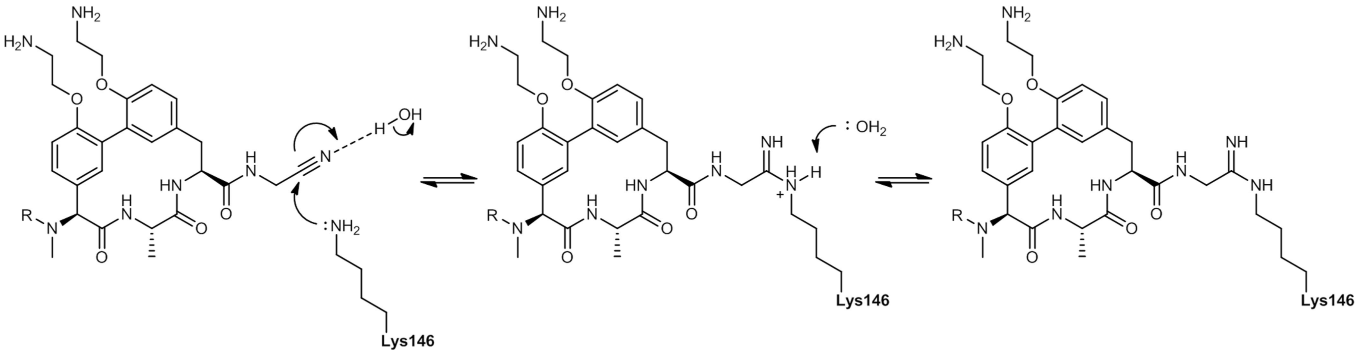
Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0483-6.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0483-6.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to P.A.S. or C.E.H.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



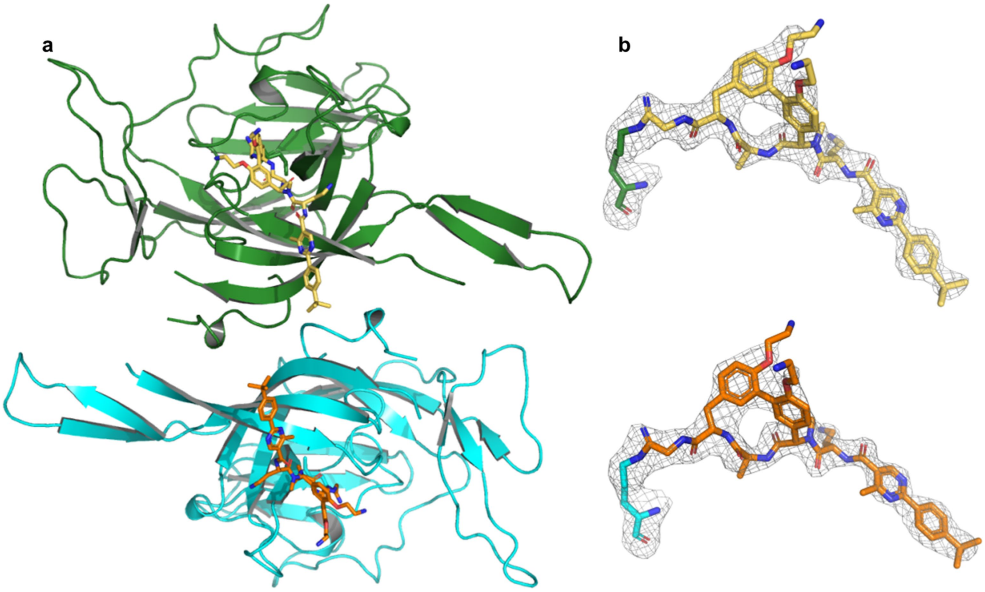
Extended Data Fig. 1 | Growth rates of *E. coli* during treatment with G0775 or when LepB expression levels are modulated. a, Time-kill data of *E. coli* ATCC 25922 when treated with G0775 at 0.25, 1, 4 and 16 times the measured MIC of 0.125 µg ml^{-1}. The data is an average of three independent experiments with error bars representing s.d. **b**, Western blot of V5-epitope-tagged LepB from an arabinose-inducible conditional LepB strain (*E. coli* UPEC CFT073) grown in the indicated sugar concentrations. This experiment was performed once. **c**, Growth-curve analysis of the same conditional LepB strain grown in the indicated sugar concentrations. Data is an average of triplicate data points with error bars representing s.d. The graph is a representative of three independent experiments.



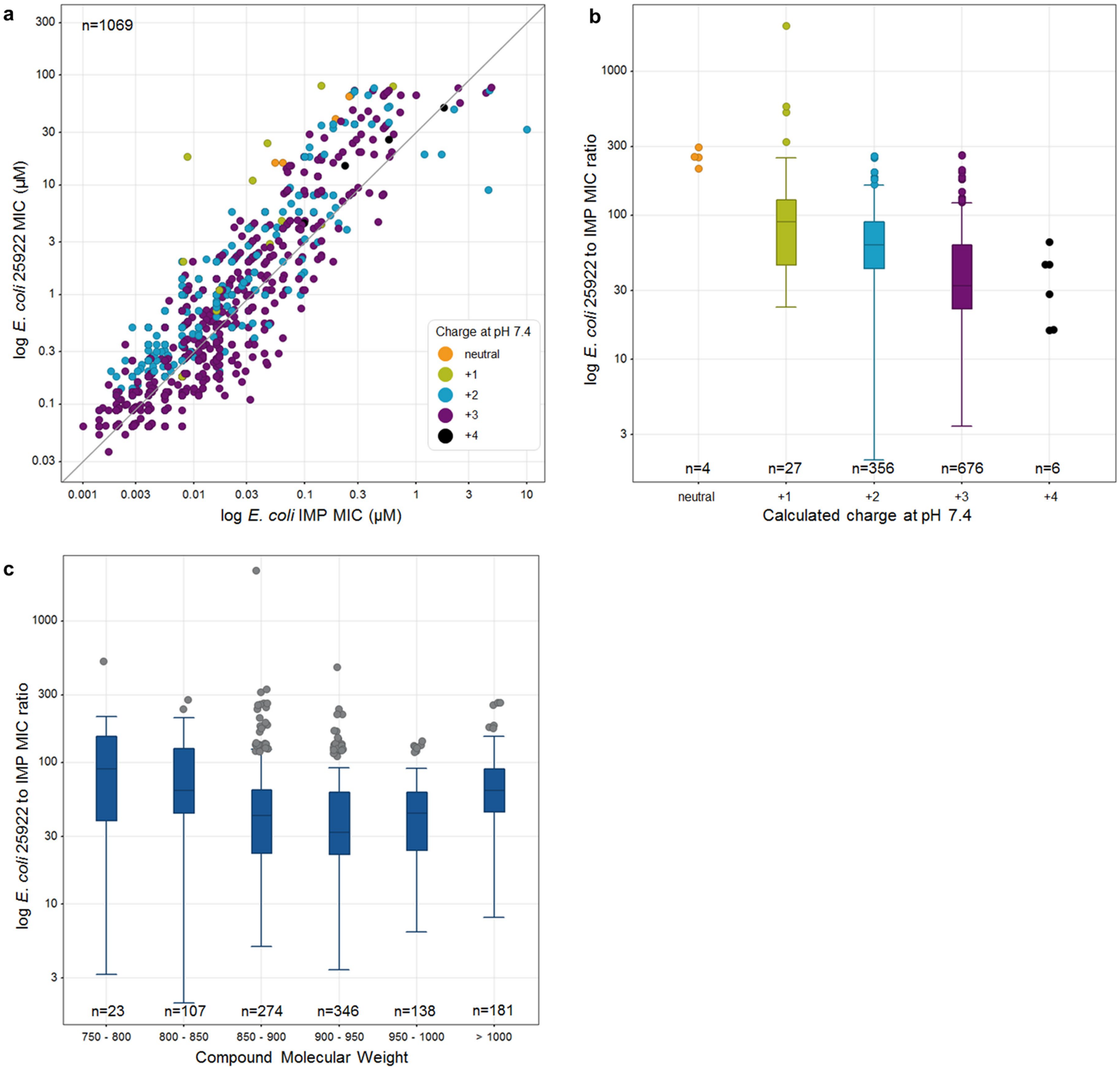
Extended Data Fig. 2 | Overlay of LepB–G0775 and LepB–arylomycin A2 (PDB ID: 1T7D). Comparison of G0775 (coloured yellow) and arylomycin (coloured pink). The catalytic lysine 146 is covalently bound to the nitrile warhead, while the serine 91 nucleophile remains unbound. For simplicity, the LepB protein has been removed from each co-structure. The comparison indicates that the macrocyclic core of G0775 maintains high similarity to the parent arylomycin macrocycle, and makes very similar interactions with the protein.


NH2
H2N
O
O
R
N
HN
H
N
O
O
O
H
N
N
H
OH
:NH2
Lys146
NH
H
H
N
+
: OH2
Lys146
NH
NH
Lys146

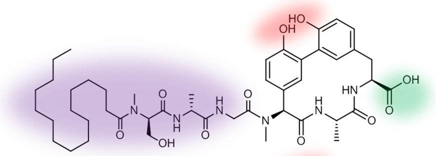
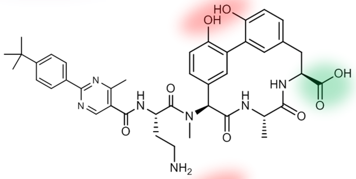
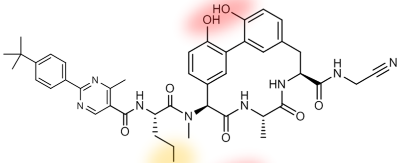
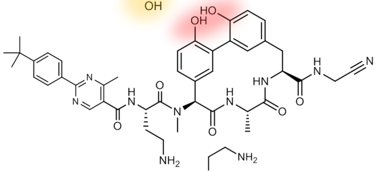
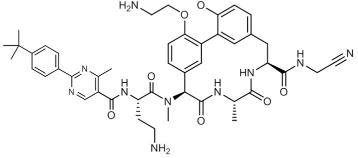
Extended Data Fig. 3 | Proposed mechanism of covalent amidine-bond formation between G0775 nitrile and lysine 146. The general base (lysine 146) involved in substrate proteolysis functions instead as a nucleophile to attack the nitrile warhead.



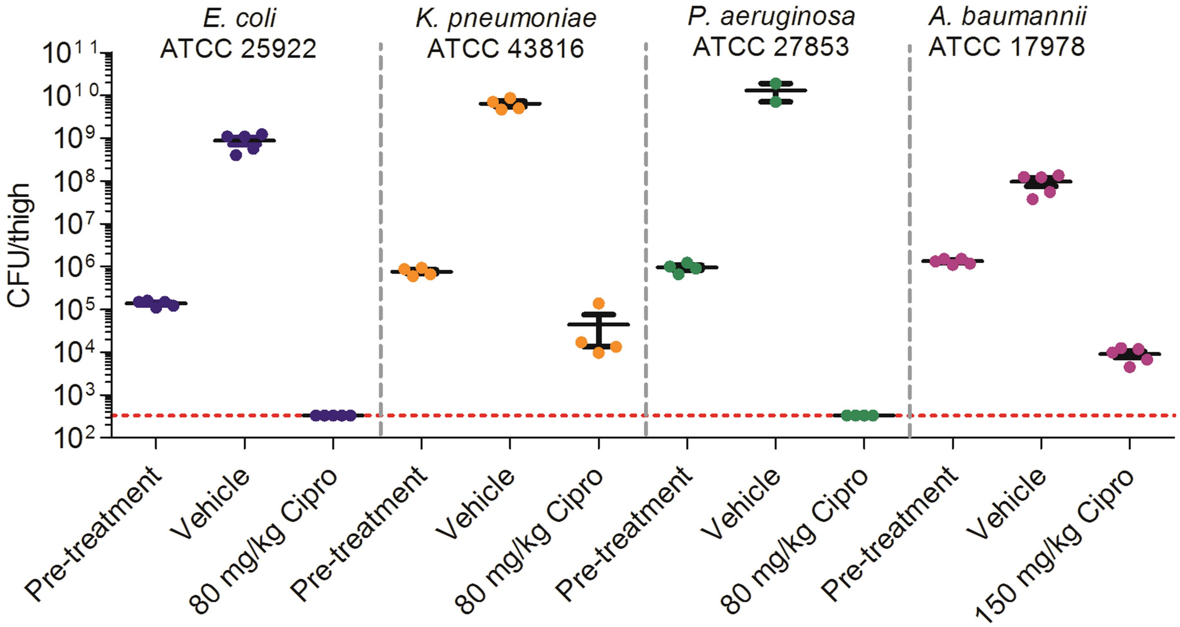
Extended Data Fig. 4 | Details of the asymmetric unit and the electron density of the LepB–G0775 complex. a, The LepB–G0775 complex crystallized in space group *I*4 with two molecules in the asymmetric unit. Chain A is coloured green, with the inhibitor shown as sticks and its carbon atoms in wheat; chain B is coloured cyan, with the inhibitor shown as sticks and its carbon atoms coloured orange. Lattice contacts across symmetry-related molecules are not shown for clarity. **b**, $mF_o - DF_c$ omit maps for each inhibitor molecule in the asymmetric unit are shown as grey mesh contoured at 3σ.



Extended Data Fig. 5 | Analysis of Gram-negative outer membrane permeability for 1,069 arylomycin analogues. a, MIC values of test compounds against *E. coli* ATCC 25922 and the outer-membrane-permeabilized *E. coli* IMP strain in the K-12 background show a consistent shift across nearly three orders of magnitude of activity. A grey line indicates a 30-fold shift in activity, and the points are coloured by the charge calculated to be the predominant species in solution at pH 7.4. Charges were calculated using MoKa 2.6.6. MIC values were measured in duplicate at a minimum, and the geometric mean was plotted. **b**, Box-and-whisker plot illustrating the IMP:25922 ratio of MIC values, for each indicated charge state from neutral to +4 charged compounds. **c**, Box-and-whisker plots showing the IMP:25922 ratio of MIC values with the compounds subdivided by molecular weight. For **b** and **c**, the central horizontal line indicates the median value for each group, and the box ends indicate the first and third quartiles in each group. The whiskers correspond to the highest non-outlier points in each group. Outliers are defined as the points outside the median value ±1.5-fold the difference between the third quartile and first quartile. Outlier points are plotted individually.

	Gram	*negative*	*negative*	*negative*	*negative*	*positive*	*positive*	*positive*	*positive*
	outer membrane	permeable	intact	intact	intact	NA	NA	NA	NA
	Species	*E. coli*	*E. coli*	*A. bau*	*P. aer*	*S. epid*	*S. aure*	*S. pneu*	*E. faec*
	Strain	K12 lptD(4213)	ATCC 25922	ATCC 17978	ATCC 27853	RP62A	USA300	ATCC 49619	ATCC 29212
	Arylomycin A-C16	16	>64	>128	> 64	0.25	32	> 64	> 64
	G3031	0.125	8	>64	>64	>64	4	>64	>64
	G8126	0.25	32	>64	>64	4	2	>64	>64
	G6850	0.03	4	8	32	2	0.5	>64	64
	G0775	0.008	0.125	1	2	0.25	0.06	64	64

Extended Data Fig. 6 | Chemical structures and biological activities of molecules along the optimization pathway from arylomycin A-C_{16} to G0775. MICs reported are in μg ml^{-1} units.



Extended Data Fig. 7 | In vivo efficacy of ciprofloxacin. Thigh infections initiated in neutropenic mice with the indicated Gram-negative bacterial species were treated with ciprofloxacin (Cipro) (*n* = 5 mice for *E. coli* and *A. baumannii*; *n* = 4 mice for *K. pneumoniae* and *P. aeruginosa*) or vehicle (*n* = 2 mice for *P. aeruginosa*, *n* = 4 mice for *K. pneumoniae*, *n* = 5 mice for *E. coli* and *A. baumannii*), and bacterial burden was quantified 20 h after infection. Ciprofloxacin was delivered subcutaneously once during the infection period, 2 h post-infection at the indicated dose. The pre-treatment and vehicle groups are identical to those shown in Fig. 4a, as these data were generated in the same studies. Broken red lines represent the limit of bacterial CFU determination, and the centre measure represents the mean with error bars representing s.e.m.

Extended Data Table 1 | MIC values of G0775 measured against *E. coli* K-12 MG1655, uropathogenic *E. coli* (UPEC) and isogenic matched strains with the indicated genetic or pharmacological manipulation

Strain	Phenotype	G0775 MIC (μg/ml)
MG1655 WT	WT	0.125
MG1655 ΔacrB	Efflux null	0.125
MG1655 ΔtolC	Efflux null	0.063
MG1655 ΔompC	Porin deficient	0.125
MG1655 ΔompF	Porin deficient	0.125
MG1655 ΔompC, ΔompF	Doubly porin deficient	0.125
MG1655 LptD4213 (imp)	Outer-membrane permeable	0.004
MG1655 WT + 4mM EDTA	Outer-membrane permeable	0.002
MG1655 WT + 50% serum	Plasma protein binding	0.125
MG1655 WT + 1% surfactant	Lung inactivation	0.125
UPEC WT	WT	0.125
UPEC – LepB Low	Reduced LepB expression level	0.008
UPEC – LepB High	Increase LepB expression level	1

WT, wild type.

Extended Data Table 2 | MIC values of G0775 measured against *E. coli* ATCC 25922 mutants that were generated spontaneously in response to plating on G0775 at 4 × MIC

Background	Mutation	LepB residue, *E. coli* numbering	Fold MIC Increase
E. coli ATCC 25922	E83K	83	16
E. coli ATCC 25922	F85S	85	8
E. coli ATCC 25922	P88L	88	16
E. coli ATCC 25922	P88Q	88	4
E. coli ATCC 25922	P88R	88	16
E. coli ATCC 25922	P88T	88	4
E. coli ATCC 25922	V133F	133	16
E. coli ATCC 25922	L142R	142	8
E. coli ATCC 25922	D143Y	143	8
E. coli ATCC 25922	V149V-KRAV	149	128
E. coli ATCC 25922	A297V	297	4
E. coli ATCC 25922	AcrB (N282Y)	-	4

Extended Data Table 3 | Cytotoxicity profile of G0775 against four different cultured mammalian cell lines

Cell Line	IC_{50} (μM)	Maximum % Inhibition at 50 μM
A549	>50	1.2
HEK-293T	>50	0.65
Jurkat	>50	-0.1
H23	>50	1.7

Cellular viability was measured after 3-day incubation of G0775 with each of the different cells by quantifying ATP levels as described in Supplementary Methods. The highest testing concentration was 50 μM.

https://doi.org/10.1038/s41586-018-0482-7

Coupling of bone resorption and formation by RANKL reverse signalling

Yuki Ikebuchi[1,7], Shigeki Aoki[1,7], Masashi Honma[1,7]*, Madoka Hayashi[1], Yasutaka Sugamori[2], Masud Khan[2], Yoshiaki Kariya[1], Genki Kato[3], Yasuhiko Tabata[4], Josef M. Penninger[5], Nobuyuki Udagawa[6], Kazuhiro Aoki[2] & Hiroshi Suzuki[1]

Receptor activator of nuclear factor-kappa B (RANK) ligand (RANKL) binds RANK on the surface of osteoclast precursors to trigger osteoclastogenesis. Recent studies have indicated that osteocytic RANKL has an important role in osteoclastogenesis during bone remodelling; however, the role of osteoblastic RANKL remains unclear. Here we show that vesicular RANK, which is secreted from the maturing osteoclasts, binds osteoblastic RANKL and promotes bone formation by triggering RANKL reverse signalling, which activates Runt-related transcription factor 2 (Runx2). The proline-rich motif in the RANKL cytoplasmic tail is required for reverse signalling, and a RANKL(Pro29Ala) point mutation reduces activation of the reverse signalling pathway. The coupling of bone resorption and formation is disrupted in RANKL(Pro29Ala) mutant mice, indicating that osteoblastic RANKL functions as a coupling signal acceptor that recognizes vesicular RANK. RANKL reverse signalling is therefore a potential pharmacological target for avoiding the reduced bone formation associated with inhibition of osteoclastogenesis.

Bone homeostasis is maintained by coordinated cycles of bone resorption and formation. The coupling signals, transmitted by osteoclasts to osteoblasts, drive the transition from bone resorption to formation[1,2]. RANKL is a transmembrane protein from the tumour necrosis factor (TNF) superfamily, and has a central role in osteoclastogenesis[3]. Previously, osteoblasts were thought to be the predominant providers of RANKL to osteoclasts[4]; however, cumulative evidence indicates that osteocytes are the major sources of RANKL during the bone remodelling process[5,6]. The exact physiological roles of osteoblastic RANKL are unknown.

Some TNF superfamily members are known to act as bidirectional signalling molecules that generate intracellular reverse signalling[7]. The RANKL-binding peptides WP9QY and OP3-4 can stimulate osteogenic activity in osteoblasts[8]. Therefore, osteoblastic RANKL potentially acts as a physiological osteogenic signal acceptor. If it does have this role, the candidate ligand for osteoblastic RANKL is RANK, or the other RANKL binding partner, osteoprotegerin (OPG)[9,10]. OPG, which is expressed in the osteoblastic lineage, is a negative regulator of RANKL forward signalling; osteoblast osteogenic activity is not reduced in OPG-deficient mice[11]. Therefore, OPG is unlikely to act as an osteogenic ligand in RANKL reverse signalling. RANK, however, is a transmembrane protein that is expressed in the osteoclastic lineage[9]. Recent reports suggest that osteoblasts that are not in direct contact with osteoclasts at the resorption pit contribute to new bone synthesis during bone remodelling[12]. A lack of cell–cell interactions between osteoblasts and osteoclasts during the bone remodelling process is also implied by the minor contribution of osteoblasts to physiological RANKL provision[5,6]. In addition, osteoclast-derived semaphorin 4D inhibits osteoblast migration to the resorption pit[13], suggesting that osteoclasts cannot stimulate RANKL reverse signalling in osteoblasts through direct interactions. However, maturing osteoclasts secrete small extracellular vesicles (SEVs) that contain RANK[14]. Therefore, it is possible that RANKL reverse signalling in osteoblasts is activated by vesicular RANK, which is secreted from osteoclasts. We tested this hypothesis in the present study.

Vesicular RANK stimulates osteoblast differentiation

First, we analysed RANK secretion by osteoclasts during their maturation. We stimulated mouse osteoclast precursor-like RAW 264.7 cells with the ectodomain of RANKL fused to glutathione *S*-transferase (GST–sRANKL). We fractionated culture supernatants by serial centrifugation and measured the content of RANK in each fraction (Fig. 1a). The pellet obtained after ultracentrifugation of the culture supernatant of maturing osteoclasts (mOCs, 60 to 120 h after GST–sRANKL stimulation) contained a higher level of RANK compared to the corresponding pellet from pre-osteoclasts prior to multinucleation (pOCs, 0 to 48 h after stimulation with GST–sRANKL). RANK was not detected in either supernatant after ultracentrifugation at 100,000*g*, indicating that the levels of soluble RANK were negligible. Co-precipitation of RANK with SEV marker proteins, which was abolished by treatment with detergents, confirmed that transmembrane RANK was incorporated into SEVs from mOCs (mOC-SEVs) (Extended Data Fig. 1). Next, we analysed the effects of mOC-SEVs on osteoblasts. We stimulated mouse osteoblast-like ST2 cells with mOC-SEVs and measured levels of osteoblast marker mRNA using quantitative PCR with reverse transcription (RT–qPCR) (Fig. 1b). Relatively rapid responses were observed for Runt-related transcription factor 2 (*Runx2*) and *Osterix* (also known as *Sp7*), both of which are central regulators of early osteoblast differentiation[15]. Early-stage differentiation markers, including type I collagen (*Col1a1*), were also upregulated. In addition, masking vesicular RANK by pre-incubation of the SEVs with GST–sRANKL abolished their stimulatory effect. These responses were confirmed in mouse primary osteoblasts (Extended Data Fig. 2). These results indicate that stimulating osteoblasts with mOC-SEVs induced Runx2 activation, and that vesicular RANK was responsible for the signal input. To further confirm the involvement of RANK in osteoblast

---

[1] Department of Pharmacy, the University of Tokyo Hospital, Faculty of Medicine, the University of Tokyo, Tokyo, Japan. [2] Department of Basic Oral Health Engineering, Graduate School of Medical and Dental Sciences, Tokyo Medical and Dental University, Tokyo, Japan. [3] Department of Bio-Matrix (Pharmacology), Graduate School of Medical and Dental Sciences, Tokyo Medical and Dental University, Tokyo, Japan. [4] Laboratory of Biomaterials, Department of Regeneration Science and Engineering, Institute for Frontier Life and Medical Sciences, Kyoto University, Kyoto, Japan. [5] Institute of Molecular Biotechnology of the Austrian Academy of Sciences, Vienna Biocenter, Vienna, Austria. [6] Department of Biochemistry, Matsumoto Dental University, Shiojiri, Japan. [7] These authors contributed equally: Yuki Ikebuchi, Shigeki Aoki, Masashi Honma. *e-mail: mhonma-tky@umin.ac.jp

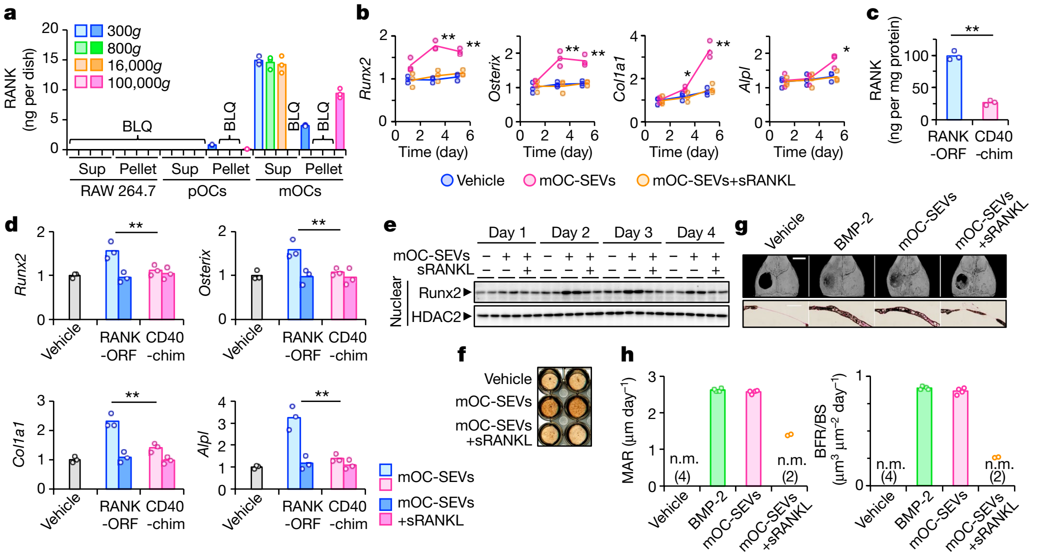


Fig. 1 | Osteogenic potency of vesicular RANK. **a**, RANK levels in fractions of supernatants from cultures of RAW 264.7 cells ($n = 3$ experiments). BLQ, below the limit of quantification. **b**, Osteoblast marker expression in ST2 cells stimulated with mOC-SEVs, mOC-SEVs pre-treated with GST–sRANKL (mOC-SEVs plus sRANKL) ($n = 3$ experiments). **c**, RANK content in SEVs from mOCs treated with RANK shRNA plus CD40-chimaera (CD40-chim) or RANK shRNA plus shRNA-resistant RANK coding region (RANK-ORF) ($n = 3$). **d**, Changes in mRNA expression levels of osteoblast markers in ST2 cells stimulated with RANK-suppressed SEVs ($n = 3$). **e**, **f**, Responses of ST2 cells stimulated with mOC-SEVs, mOC-SEVs plus sRANKL, or vehicle. Time course of nuclear Runx2 (**e**) and von Kossa-stained images (**f**). **g**, **h**, Mouse calvarial defect model treated with mOC-SEVs, mOC-SEVs plus sRANKL, BMP-2, or vehicle. Representative images of microfocal computed tomography of whole calvaria and von Kossa-stained calvaria sections out of 4 mice in each treated group (**g**); mineral apposition rate (MAR) and bone formation rate (BFR/BS) in the defect region (**h**) ($n = 4$; n.m., not measurable). **a**–**d**, **h**, Data are mean, with individual points representing biologically independent samples or mice. Unpaired two-sided Student's *t*-test (**c**) or one-way ANOVA followed by Student–Newman–Keuls test (**b**, **d**); $^{*}P < 0.05$, $^{**}P < 0.01$.

stimulation, we prepared SEVs from mOCs under conditions of RANK suppression. We used a chimeric CD40 construct (CD40 chimaera[16]), in which the cytoplasmic tail was replaced with RANK, and introduced it into RAW 264.7 cells along with short hairpin RNA (shRNA) against RANK. We then stimulated osteoclast maturation using a CD40 antibody. SEVs prepared under these conditions contained approximately 20% of the RANK found in SEVs prepared under control conditions (Fig. 1c). Stimulating ST2 cells with these CD40-chimaera SEVs caused reduced stimulation of osteoblast marker expression in comparison to the control SEVs (Fig. 1d). These results indicate that RANK is a major determinant of the osteoblast-stimulating effect of mOC-SEVs. We also confirmed that Runx2 was activated in ST2 cells stimulated with mOC-SEVs. Immunoblotting revealed that, in the presence of mOC-SEVs, levels of Runx2 in the nucleus were increased and reached a maximum at day 3. RANK-masking pre-treatment of SEVs inhibited their stimulatory effect (Fig. 1e). Transient activation of Runx2 would be expected to promote osteoblastic mineralization. Therefore, we cultured ST2 cells for 6 days in the presence of mOC-SEVs, and performed von Kossa staining after an additional 12 days of culture in the absence of mOC-SEVs (Fig. 1f). The presence of mOC-SEVs increased osteoblastic mineralization; this effect was inhibited by RANK-masking pre-treatment of the mOC-SEVs. Finally, we examined the effect of mOC-SEVs on bone formation in vivo. We performed trepanation on mouse calvaria, and placed a gelatin hydrogel sheet impregnated with mOC-SEVs on the injury. Four weeks after the procedure, mice treated with mOC-SEVs exhibited marked healing of the skull injury, whereas mice treated with the control gelatin hydrogel showed little or no healing (Fig. 1g). RANK-masking pre-treatment prevented this stimulatory effect. Histomorphometric analyses confirmed the upregulation of bone formation in the presence of mOC-SEVs (Fig. 1h).

RANKL-crosslinking activates reverse signalling

Next, we examined whether osteoblastic RANKL was the cellular surface acceptor for vesicular RANK. Primary osteoblasts derived from RANKL-deficient mice were cultured in the presence of mOC-SEVs for 6 days, and von Kossa staining was performed after an additional 15 days of culture (Fig. 2a). The promotion of mineralization by mOC-SEVs in wild-type osteoblasts was almost completely abolished in RANKL-deficient osteoblasts. In addition, induction of osteoblast markers by mOC-SEVs was also abolished in RANKL-deficient osteoblasts (Fig. 2b). This suggests that osteoblastic RANKL acts as an acceptor for vesicular RANK. We also confirmed that the addition of recombinant OPG inhibited the effect of mOC-SEVs on osteoblastic mineralization and induction of osteoblast markers (Extended Data Fig. 3), supporting the above observations. We then explored the intracellular signalling pathways involved in Runx2 activation. On the basis of reports showing that mechanistic target of rapamycin complex 1 (mTORC1) positively regulates osteoblast differentiation via activation of Runx2[17], we examined the effect of the mTORC1 inhibitor rapamycin (Extended Data Fig. 4). Rapamycin diminished levels of nuclear Runx2 following stimulation with mOC-SEVs. We therefore examined the activation of mTORC1 and its major upstream signalling molecules PI3K and Akt in ST2 cells treated with mOC-SEVs (Fig. 2c). Stimulation with mOC-SEVs resulted in sustained activation of PI3K, Akt and mTORC1; RANK-masking pre-treatment of mOC-SEVs greatly reduced the activation of PI3K–Akt–mTORC1 pathway. Next, we examined the triggering mechanism of RANKL reverse signalling. Simply adding recombinant RANK or OPG recombinant proteins did not stimulate osteoblastic mineralization (Extended Data Figs. 3, 5), suggesting that multimeric assembly of RANKL by vesicular RANK may be necessary to trigger signalling. To examine this, it is helpful to use a RANKL-crosslinking agent other than vesicular RANK; one such agent is the anti-RANKL single-chain variable fragment (scFv) fused with an isoleucine zipper, which forms a homotrimer. We screened phage libraries displaying scFv against the RANKL ectodomain, and rearranged selected clones into trimeric forms. We selected a clone (αR-tri) that produced a sufficient yield of recombinant protein, and expressed and purified the protein using affinity chromatography

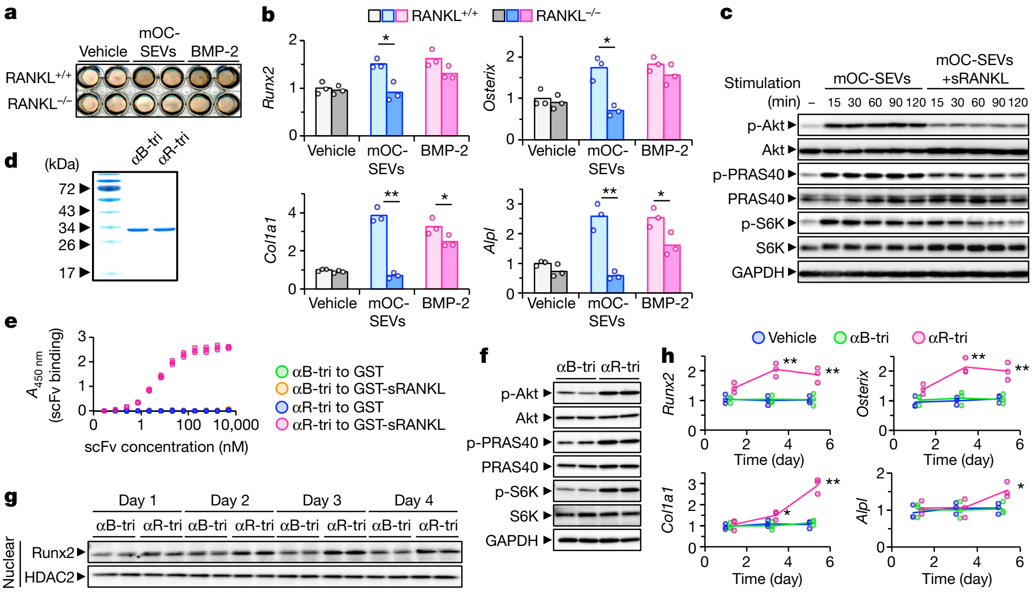


Fig. 2 | Activation of reverse signalling by RANKL crosslinking. **a**, **b**, Responses of osteoblasts, isolated from RANKL-deficient mice or wild-type littermates, and stimulated with mOC-SEVs, BMP-2, or vehicle; von Kossa-stained images (**a**) and changes in mRNA expression levels (**b**) ($n = 3$). **c**, Activation of PI3K–Akt–mTORC1 pathway in ST2 cells stimulated with mOC-SEVs, mOC-SEVs plus sRANKL, or vehicle. **d**, SDS-PAGE of αR-tri and anti-BSA trimer control (αB-tri) used for analyses. **e**, Binding of αR-tri to the RANKL ectodomain ($n = 3$ biologically independent samples). **f**, Activation of PI3K–Akt–mTORC1 pathway in ST2 cells stimulated with αR-tri or αB-tri. **g**, Nuclear Runx2 protein in ST2 cells following stimulation with αR-tri or αB-tri. **h**, Expression of osteoblast markers in ST2 cells stimulated with αR-tri or αB-tri ($n = 3$). In **b**, **h**, data are mean, with individual points representing biologically independent samples. One-way ANOVA followed by Student–Newman–Keuls test; $^*P < 0.05$, $^{**}P < 0.01$.

(Fig. 2d). The binding affinity of αR-tri to RANKL ectodomain was approximately 5 nM (Fig. 2e). Immunoblot analyses of ST2 cells treated with αR-tri for 20 min confirmed that this treatment activated the PI3K–Akt–mTORC1 pathway (Fig. 2f). In addition, nuclear Runx2 was upregulated by αR-tri treatment and reached a maximum 72 h after stimulation (Fig. 2g). This was consistent with the observations obtained using vesicular RANK. Furthermore, the increase in nuclear Runx2 following stimulation was diminished by treatment with rapamycin (Extended Data Fig. 5). We also assessed the influence of αR-tri on osteoblast marker mRNA in ST2 cells (Fig. 2h). The timings of the increase in osteoblast markers were similar to those obtained from stimulation with mOC-SEVs. These results were confirmed in mouse primary osteoblasts (Extended Data Fig. 5).

RANKL reverse signalling exerts biphasic influence

Previous reports have shown that Runx2 promotes early-stage osteoblast differentiation but suppresses late-stage differentiation[15]. We therefore assumed that RANKL reverse signalling has a biphasic influence on osteoblast differentiation. Stimulation of ST2 cells with mOC-SEVs or αR-tri from day 1 to 6 after cell seeding promoted early expression of *Col1a1*; osteocalcin expression at the late stage was also accelerated (Fig. 3a). Conversely, stimulation with mOC-SEVs or αR-tri from day 13 to 18 suppressed the induction of osteocalcin (Fig. 3b). In addition, von Kossa staining of ST2 cells treated in the same manner revealed that stimulation at the early stage promoted osteoblastic mineralization, whereas stimulation at the late stage suppressed osteoblastic mineralization (Fig. 3c). These results were confirmed in mouse primary osteoblasts (Extended Data Fig. 6). We also examined whether the effects of RANKL reverse signalling on osteoblasts are attributable mostly to the activation of Runx2. We introduced a dominant-negative form of Runx2 (DN-Runx2) into ST2 cells, followed by stimulation with αR-tri from day 1 to 6. The increase in early-stage *Col1a1* induction and late-stage osteocalcin induction by αR-tri was nearly completely suppressed by DN-Runx2 (Fig. 3d). Likewise, we introduced DN-Runx2 into ST2 cells at day 12, and stimulated the cells with αR-tri from day 13 to 18. In mock-infected ST2 cells, the induction of osteocalcin expression was strongly suppressed; this suppressive effect was abolished by DN-Runx2 (Fig. 3e). The results of von Kossa staining (Fig. 3f) were consistent with these results.

Osteoblastic RANKL is a coupling signal acceptor

The timing of vesicular-RANK secretion from osteoclasts and the biphasic properties of RANKL reverse signalling suggest that the reverse signalling may contribute to coupling signals. To test this in vivo, we first evaluated the effects of suppressing SEVs. On the basis of previous reports[18], we used the neutral sphingomyelinase inhibitor GW4869 to disrupt the secretion of SEVs. We confirmed in vitro that GW4869 reduced the secretion of vesicular RANK from mOCs, as well as moderately affecting osteoclast functions (Extended Data Fig. 7). Then, we evaluated the effect of GW4869 on the in vivo coupling of bone resorption and formation. We administered GST–sRANKL to mice intraperitoneally to induce a transient increase in osteoclastogenesis and the coupled upregulation of bone formation[19]. The influence of co-administering GW4869 was evaluated by histomorphometric analyses (Fig. 4a). GW4869 significantly suppressed the upregulation of bone formation; this was associated with a transient increase in osteoclast maturation. We also examined the effect of a neutralizing antibody on mOC-SEVs. In this scenario, the antibody recognizes a specific molecule presented on the surface of mOC-SEVs and sterically hinders the interaction of vesicular RANK with osteoblastic RANKL. We performed a proteomics analysis of mOC-SEVs and identified immunoglobulin superfamily member 8 (IGSF8) as a candidate target protein (Supplementary Table 1). We then screened a phage display library and obtained two different scFv clones that showed high-affinity binding to the IGSF8 ectodomain (Extended Data Fig. 7). The scFv recombinant proteins in the trimeric form did not affect osteoclast functions, whereas pre-treatment of mOC-SEVs with anti-IGSF8 scFv trimers (αI-tri1/2) efficiently abolished the stimulatory effects of mOC-SEVs on osteoblasts (Extended Data Fig. 7). On the basis of these results, we evaluated the effect of αI-tri1/2 on the in vivo coupling process using the same mouse model as described above (Fig. 4b). The upregulation of bone formation was significantly suppressed by the co-administration of αI-tri1/2. These results indicate that mOC-SEVs were involved in the coupling of bone resorption and formation. Next, we examined

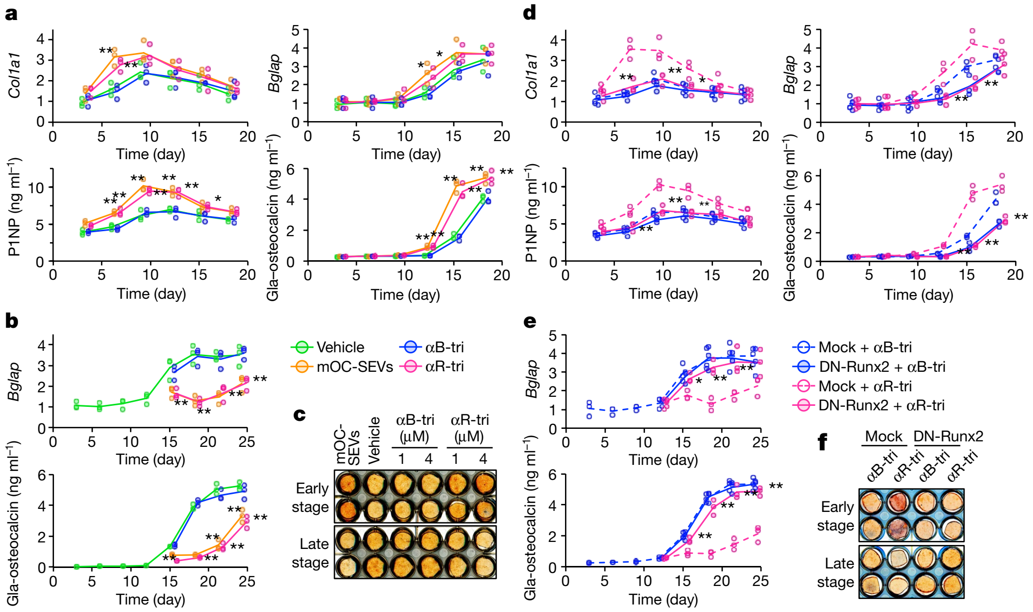


Fig. 3 | Biphasic effect of RANKL reverse signalling. a, b, Time courses of early-osteoblastic markers (*Col1a1* mRNA and procollagen type 1 N-terminal propeptide (P1NP) reflect the expression of type I collagen) and late-osteoblastic markers (osteocalcin mRNA (*Bglap*) and γ-carboxyglutamic acid (Gla)−osteocalcin protein) in ST2 cells stimulated during early stage (**a**) or during late stage (**b**) with mOC-SEVs, αR-tri, αB-tri, or vehicle ($n = 3$). **c**, Von Kossa staining of ST2 cells stimulated with mOC-SEVs or αR-tri during the early or late stage of differentiation. **d**, **e**, Effect of dominant negative Runx2 (DN-Runx2) on time course of osteoblastic marker expression in ST2 cells stimulated during early stage (**d**) or late stage (**e**) with αR-tri or αB-tri ($n = 3$). **f**, Von Kossa staining of ST2 cells expressing DN-Runx2 before stimulation with αR-tri or αB-tri. In **a**, **b**, **d**, **e**, lines represent the mean of three biologically independent samples. One-way ANOVA followed by Student–Newman–Keuls test; *$P < 0.05$, **$P < 0.01$.

the involvement of RANKL reverse signalling. To establish a mouse model in which RANKL reverse signalling is selectively suppressed, we analysed the roles of a characteristic proline-rich motif (PRM) in the RANKL cytoplasmic tail[20]. Because PRMs interact with Src homology 3 (SH3) domains[21], we focused on Src family kinases (SFKs), which contain SH3 domains and activate PI3K[22]. SFK inhibitors and dominant-negative form of Src exhibited a suppressive effect on RANKL reverse signalling (Extended Data Fig. 8). Furthermore, overexpression of a RANKL point mutant, in which a proline residue in the PRM was mutated to alanine (Pro29Ala or Pro39Ala), reduced mOC-SEV-induced activation of PI3K (Extended Data Fig. 8); this indicates that RANKL PRM is required for the activation of downstream signalling. Therefore, we introduced a point mutation into exon 1 of the *RANKL* gene (also known as *Tnfsf11*) using a Cre–loxP system. There was little difference in RANKL expression or surface presentation between osteoblasts derived from wild-type mice and those derived from the homozygous mutant mice (hereafter referred to as *RANKL*^{P29A} mice) (Extended Data Fig. 8). However, the *RANKL*^{P29A} mutation resulted in reduced PI3K and mTORC1 activation in osteoblasts stimulated with mOC-SEVs (Fig. 4c), indicating that RANKL

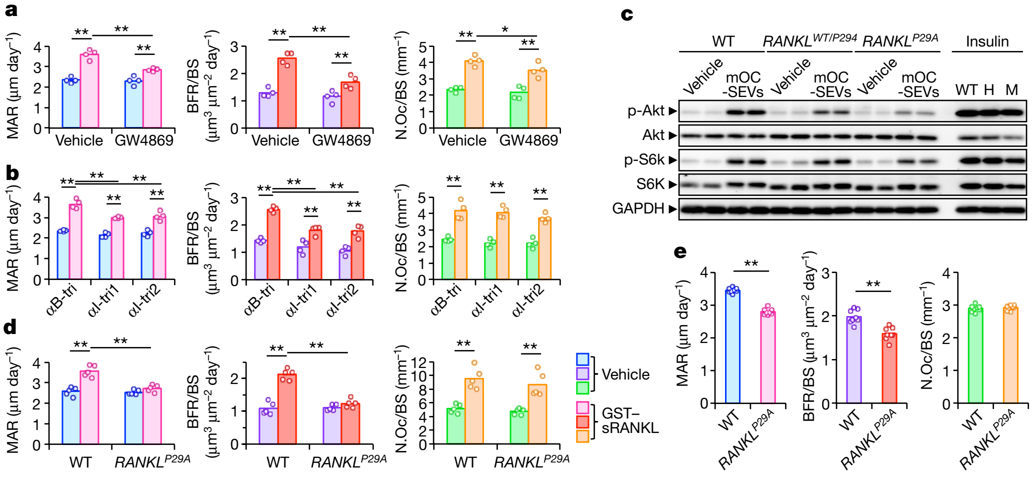


Fig. 4 | Coupling of bone resorption and formation by RANKL reverse signalling. a, b, The influence of counteractive agents against mOC-SEVs, GW4869 (**a**) or αI-tri1/2 (**b**), on in vivo coupling of bone resorption and formation; the number of osteoclasts (N.Oc/BS), bone formation rate and mineral apposition rate following intraperitoneal administration of GST–sRANKL or vehicle. ($n = 4$). **c**, Activation of PI3K and mTORC1 in primary osteoblasts isolated from *RANKL*^{P29A} mice or wild-type littermates. Insulin was used as positive control. **d**, Evaluation of the coupling of bone resorption and formation in *RANKL*^{P29A} mice and wild-type littermates, administered intraperitoneally with GST–sRANKL ($n = 5$). **e**, The influence of *RANKL*^{P29A} mutation on high bone turnover phenotypes in *OPG*$^{+/-}$ mice ($n = 8$). **a**, **b**, **d**, **e**, Data are mean with individual scatter plots, representing individual mice. One-way ANOVA followed by Student–Newman–Keuls test (**a**, **b**, **d**.) or unpaired two-sided Student’s *t*-test (**e**); *$P < 0.05$, **$P < 0.01$.

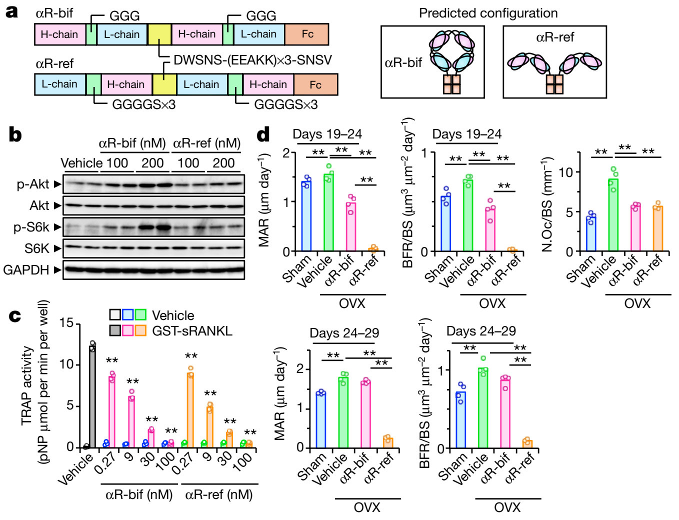


Fig. 5 | Pharmacological effect of activating RANKL reverse signalling. **a**, Structures of αR-bif and αR-ref constructs. **b**, Activation of PI3K and mTORC1 in ST2 cells stimulated with αR-bif or αR-ref. **c**, Inhibition of osteoclast maturation by αR-bif and αR-ref ($n = 3$ (vehicle), 4 (GST–sRANKL)). **d**, Pharmacological effects of αR-bif and αR-ref in OVX mice; number of osteoclasts, bone formation rate and mineral apposition rate in OVX mice following administration of αR-bif ($n = 4$), αR-ref ($n = 3$), or vehicle ($n = 4$), or sham-operated mice with vehicle administration ($n = 4$). In **c**, **d**, data are mean, with individual points representing biologically independent samples or mice. One-way ANOVA followed by Student–Newman–Keuls test; $^{**}P < 0.01$.

reverse signalling was suppressed in osteoblasts from *RANKL*^{P29A} mice. On the basis of these results, we evaluated the influence of the *RANKL*^{P29A} mutation on the in vivo coupling process. We administered GST–sRANKL to *RANKL*^{P29A} mice and wild-type littermates, and performed histomorphometric analyses (Fig. 4d). Osteoclast maturation was increased in both the *RANKL*^{P29A} and wild-type littermates. By contrast, upregulation of bone formation was strongly suppressed in the *RANKL*^{P29A} mutants in comparison to wild-type littermates. In this mouse model, osteocytic RANKL showed a negligible contribution to transient osteoclastogenesis[19]; the roles of osteoblastic RANKL reverse signalling could therefore be evaluated separately. The above results indicate that osteoblastic RANKL reverse signalling contributes to the coupling of bone resorption and formation. To confirm this, we investigated the influence of the *RANKL*^{P29A} mutation on high bone turnover phenotypes in *OPG* (also known as *Tnfrsf11b*) heterozygous (*OPG*$^{+/-}$) mice. We performed histomorphometric analyses of lumbar spine in *RANKL*^{P29A};*OPG*$^{+/-}$ mice and RANKL wild-type (*RANKL*^{WT};*OPG*$^{+/-}$) littermates (Fig. 4e). *RANKL*^{P29A};*OPG*$^{+/-}$ mice exhibited reduced bone formation parameters in comparison to *RANKL*^{WT};*OPG*$^{+/-}$ littermates, whereas osteoclast number at the bone surface was not affected by the *RANKL*^{P29A} mutation. These results indicate that suppression of RANKL reverse signalling leads to inefficiency of bone formation under physiological conditions. The influence of the *RANKL*^{P29A} mutation on bone phenotypes in 17-week-old *OPG*$^{+/+}$ mice was consistent with these observations (Extended Data Fig. 9).

RANKL reverse signalling as a pharmacological target

In light of the bifunctional properties of RANKL, we examined the possibility of designing a biological agent to simultaneously activate RANKL reverse signalling in osteoblasts and inhibit RANKL forward signalling to osteoclast precursors. To achieve an acceptable in vivo half-life and RANKL-crosslinking property, we searched for suitable linker peptide sequences and scFv orientations to rearrange a RANKL-specific scFv to produce IgG1–Fc fusion proteins with four RANKL-binding sites (Fig. 5a). We identified a bifunctional construct that both inhibited forward signalling and activated reverse signalling (αR-bif), and a second reference construct that exhibited only the inhibitory effect on RANKL forward signalling (αR-ref). We observed activation of RANKL reverse signalling in ST2 cells stimulated with αR-bif, but not in cells treated with αR-ref (Fig. 5b). Conversely, both αR-bif and αR-ref had inhibitory effects on osteoclast maturation in vitro (Fig. 5c). On the basis of these results, we evaluated the effects of these constructs in an ovariectomized (OVX) mouse model (Fig. 5d). OVX resulted in the upregulation of bone formation, which was coupled with increased osteoclast numbers. Treatment with αR-ref significantly decreased osteoclast numbers and bone formation in both the first (days 19–24) and second (days 24–29) half-periods. Similar to treatment with αR-ref, treatment with αR-bif also decreased osteoclast numbers, but the suppressive effect on bone formation was moderate in the first half-period and weaker in the second half-period. These results indicate that activation of RANKL reverse signalling can prevent decreased bone formation by compensating for the shortage of coupling signals. The recovery of bone formation was greater in the second half-period, consistent with the biphasic property of Runx2 activation downstream of RANKL reverse signalling.

Discussion

Among osteoclast-derived coupling-signal mediators, semaphorin 4D, which is secreted from pre-osteoclasts shortly after RANKL stimulation, inhibits osteoblast differentiation[13], whereas collagen triple helix repeat containing 1 (Cthrc1), which is secreted from bone-resorbing osteoclasts, promotes late-stage differentiation[23]. Our results indicate that RANKL reverse signalling in osteoblasts may prepare osteoblasts for further maturation promoted by Cthrc1 after the suppressive effect of semaphorin 4D is diminished. In addition, the timing of vesicular RANK secretion is similar to that of sphingosine-1-phosphate and induction of Wnt10b in osteoclasts[24], indicating that these factors cooperate to regulate osteoblast differentiation.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0482-7.

Received: 22 July 2016; Accepted: 2 August 2018;
Published online 5 September 2018.

1. Sims, N. A. & Martin, T. J. Coupling the activities of bone formation and resorption: a multitude of signals within the basic multicellular unit. *Bonekey Rep.* **3**, 481 (2014).
2. Florencio-Silva, R., Sasso, G. R., Sasso-Cerri, E., Simoes, M. J. & Cerri, P. S. Biology of bone tissue: structure, function, and factors that influence bone cells. *BioMed Res. Int.* **2015**, 421746 (2015).
3. Kong, Y. Y. et al. OPGL is a key regulator of osteoclastogenesis, lymphocyte development and lymph-node organogenesis. *Nature* **397**, 315–323 (1999).
4. Suda, T. et al. Modulation of osteoclast differentiation and function by the new members of the tumor necrosis factor receptor and ligand families. *Endocr. Rev.* **20**, 345–357 (1999).
5. Nakashima, T. et al. Evidence for osteocyte regulation of bone homeostasis through RANKL expression. *Nat. Med.* **17**, 1231–1234 (2011).
6. Xiong, J. et al. Matrix-embedded cells control osteoclast formation. *Nat. Med.* **17**, 1235–1241 (2011).
7. Kisiswa, L. et al. TNFα reverse signaling promotes sympathetic axon growth and target innervation. *Nat. Neurosci.* **16**, 865–873 (2013).
8. Sugamori, Y. et al. Peptide drugs accelerate BMP-2-induced calvarial bone regeneration and stimulate osteoblast differentiation through mTORC1 signaling. *BioEssays* **38**, 717–725 (2016).
9. Boyce, B. F. & Xing, L. The RANKL/RANK/OPG pathway. *Curr. Osteoporos. Rep.* **5**, 98–104 (2007).
10. Martin, T. J. & Sims, N. A. RANKL/OPG; Critical role in bone physiology. *Rev. Endocr. Metab. Disord.* **16**, 131–139 (2015).
11. Fei, Q. et al. Osteogenic growth peptide enhances the proliferation of bone marrow mesenchymal stem cells from osteoprotegerin-deficient mice by CDK2/cyclin A. *Acta Biochim. Biophys. Sin. (Shanghai)* **42**, 801–806 (2010).
12. Andersen, T. L. et al. Understanding coupling between bone resorption and formation: are reversal cells the missing link? *Am. J. Pathol.* **183**, 235–246 (2013).
13. Negishi-Koga, T. et al. Suppression of bone formation by osteoclastic expression of semaphorin 4D. *Nat. Med.* **17**, 1473–1480 (2011).
14. Huynh, N. et al. Characterization of regulatory extracellular vesicles from osteoclasts. *J. Dent. Res.* **95**, 673–679 (2016).
15. Komori, T. Regulation of osteoblast differentiation by transcription factors. *J. Cell. Biochem.* **99**, 1233–1239 (2006).

16. Kadono, Y. et al. Strength of TRAF6 signalling determines osteoclastogenesis. *EMBO Rep*. **6**, 171–176 (2005).
17. Singha, U. K. et al. Rapamycin inhibits osteoblast proliferation and differentiation in MC3T3–E1 cells and primary mouse bone marrow stromal cells. *J. Cell. Biochem*. **103**, 434–446 (2008).
18. Trajkovic, K. et al. Ceramide triggers budding of exosome vesicles into multivesicular endosomes. *Science* **319**, 1244–1247 (2008).
19. Tomimori, Y. et al. Evaluation of pharmaceuticals with a novel 50-hour animal model of bone loss. *J. Bone Miner. Res*. **24**, 1194–1205 (2009).
20. Kariya, Y., Honma, M., Aoki, S., Chiba, A. & Suzuki, H. Vps33a mediates RANKL storage in secretory lysosomes in osteoblastic cells. *J. Bone Miner. Res*. **24**, 1741–1752 (2009).
21. Saksela, K., Cheng, G. & Baltimore, D. Proline-rich (PxxP) motifs in HIV-1 Nef bind to SH3 domains of a subset of Src kinases and are required for the enhanced growth of Nef$^+$ viruses but not for down-regulation of CD4. *EMBO J*. **14**, 484–491 (1995).
22. Cuevas, B. D. et al. Tyrosine phosphorylation of p85 relieves its inhibitory activity on phosphatidylinositol 3-kinase. *J. Biol. Chem*. **276**, 27455–27461 (2001).
23. Takeshita, S. et al. Osteoclast-secreted CTHRC1 in the coupling of bone resorption to formation. *J. Clin. Invest*. **123**, 3914–3924 (2013).
24. Pederson, L., Ruan, M., Westendorf, J. J., Khosla, S. & Oursler, M. J. Regulation of bone formation by osteoclasts involves Wnt/BMP signaling and the chemokine sphingosine-1-phosphate. *Proc. Natl Acad. Sci. USA* **105**, 20764–20769 (2008).

Acknowledgements We thank K. Toritani, M. Maeda, and M. A. Al Mamun for their technical assistance. This work was supported in part by JSPS KAKENHI grant numbers JP16H05445, JP24390349, JP26713045, JP24890048, JP25670632; and by MEXT KAKENHI grant number JP22136015.

Reviewer information *Nature* thanks S. Takeda. M. Zaidi and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions Y.I. and S.A. performed in vitro and in vivo experiments, analysed data, and wrote the manuscript. M.Ho. designed the study, performed in vitro experiments, analysed data, and wrote and revised the manuscript. Y.S., M.K., G.K. and Y.T. designed and/or performed experiments using the calvarial defect model. M.Ha. performed construction and evaluation of biologics. Y.K. performed RANK suppression experiments. N.U. and J.M.P. designed and/or performed experiments using RANKL-deficient cells. K.A. and H.S. supervised the study and revised the manuscript.

Competing Interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0482-7.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0482-7.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to M.H.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Preparation of cells. Mouse primary osteoblasts were collected from the calvaria of 2- to 4-day-old C57BL/6 mice (Japan SLC) or RANKL-deficient mice[3] using previously reported methods[25]. RAW 264.7 cells (Dainippon Sumitomo Pharma) and ST2 cells (RIKEN Cell Bank) were maintained in α-MEM (Sigma-Aldrich) containing 10% fetal bovine serum (FBS, Biowest), 2 mM L-glutamine (Nacalai Tesque), and penicillin-streptomycin (PCSM, Life Technologies). The 293A cells (Life Technologies) were maintained in DMEM (Nacalai Tesque) containing 10% FBS, 2 mM L-glutamine, non-essential amino acids (NEAA, Life Technologies), and PCSM. The 293FT cells (Life Technologies) were maintained in DMEM containing 10% FBS, 2 mM L-glutamine, NEAA, PCSM, and G-418 (Nacalai Tesque). All cells tested negative for mycoplasma. Additional authentication of cell lines was not performed after acquisition from the distributors.

Preparation of recombinant GST–sRANKL. To obtain recombinant GST–sRANKL, an *Escherichia coli* expression vector was constructed by subcloning the ectodomain of RANKL (Pro143–Asp316) into pGEX-5X-2 (Clontech). Rosetta 2 *E. coli* were transformed with this plasmid and cultured in LB medium at 37 °C until $A_{600\,\mathrm{nm}}$ reached 0.8. Isopropylthio-β-D-galactoside (IPTG) was added for a final concentration of 20 μM, and cells were cultured for an additional 20 h at 20 °C. Cells were collected and lysed (in 50 mM phosphate buffer, pH 8.0, 150 mM NaCl, 1% Triton X-100, 10% glycerol, 1 mM EDTA, and protease inhibitors (Roche Diagnostics)) by sonication. After removal of cell debris by centrifugation, the supernatant was collected, and GST–sRANKL was purified using Glutathione Sepharose 4B gel (GE Healthcare) according to the manufacturer's protocol. After changing the eluate to phosphate buffered saline (PBS, pH 7.4) using PD-10 columns (GE Healthcare), protein concentration was determined using a Mouse RANKL Quantikine ELISA Kit (R&D Systems) according to the manufacturer's protocol. We used a conventional limulus amebocyte lysate test kit (Seikagaku) to confirm that endotoxin contamination was negligible.

Fractionation of osteoclast-secreting products. RAW 264.7 cells were seeded (1×10^4 cells/cm^2) in α-MEM containing 10% FBS, 2 mM L-glutamine, 100 ng/ml GST–sRANKL, and 10 ng/ml macrophage colony stimulating factor (M-CSF, R&D Systems). To analyse the products of osteoclast secretion, culture supernatants were subjected to serial centrifugations (300*g* for 10 min, 800*g* for 10 min, 16,000*g* for 20 min, and 100,000*g* for 60 min), and each pellet and supernatant was analysed. For further analyses, mOC-SEVs, collected by serial centrifugations, were washed and resuspended in PBS. The protein concentration of the resulting suspension was determined using a BCA Protein Assay Kit (Thermo Scientific). To mask vesicular RANK, mOC-SEVs were incubated with 10 μg/ml GST–sRANKL for 5 h at 20 °C, and unbound GST–sRANKL was removed by centrifugation at 100,000*g* for 60 min.

Quantification of RANK protein by ELISA. Multiwell plates were coated with goat RANK polyclonal antibody (R&D Systems; 2 μg/ml in 50 mM sodium carbonate buffer, pH 9.6) for 12 h at 4 °C, and 1% bovine serum albumin (BSA) in PBS was used in the subsequent blocking step for 1 h at room temperature. After three washes with washing buffer (PBS containing 0.1% Tween 20), the diluted samples were incubated for 12 h at 4 °C. After three washes, rabbit RANK polyclonal antibody (Cell Signaling Technology; 2 μg/ml) was incubated for 12 h at 4 °C. After four washes, horseradish peroxidase (HRP)-labelled donkey anti-rabbit IgG antibody (GE Healthcare; diluted 1:5000) was incubated for 1 h at room temperature. After five washes, 3,3′,5,5′-tetramethylbenzidine solution (Nacalai Tesque) was used with 1 M sulfuric acid as quencher; absorbance was measured at 450 nm with a microplate reader (Thermo Scientific). The concentration of RANK in each sample was evaluated using recombinant RANK–Fc protein (R&D Systems) as standard.

Evaluation of osteogenesis in vitro. ST2 cells or primary osteoblasts were seeded (3×10^4 cells/cm^2) in α-MEM containing 10% FBS, 2 mM L-glutamine, PCSM, and 2 mM β-glycerophosphate (Wako); the culture medium was replenished every 3 days. Osteogenic reagents, purified mOC-SEVs (50 to 75 μg protein obtained from 3 to 6×10^4 mOCs during 48 h, and containing 5 to 10 ng of RANK), or scFv trimers (1 or 4 μM) were added to the medium on days 0 to 6 to stimulate osteoblasts during early-stage differentiation; and days 12 to 18 (ST2) or days 15 to 21 (primary osteoblasts) to stimulate osteoblasts during late-stage differentiation. Osteoblasts were treated with Runx2 or DN-Runx2 using adenoviruses 24 h before exposure to αR-tri. Total RNA was extracted using RNAiso Plus (TaKaRa). The mRNA expression levels of osteoblast markers were determined by RT–qPCR using SYBR GreenER qPCR SuperMix Universal (Life Technologies) and an Eco Real-Time PCR System (Illumina) with the associated software. The expression level of each marker gene was normalized to that of mouse glyceraldehyde-3-phosphate dehydrogenase (*Gapdh*). The primer sequences used for the analyses were as follows: 5′-GAGTCAGATTACAGATCCCA-3′ and 5′-TGGCTCTTCTTACTGAGAGA-3′ for mouse *Runx2*; 5′-CCCCAAGATGTCTATAAGCC-3′ and 5′-CGCTCTAGCTCCTGACAGTT-3′ for mouse *Sp7*; 5′-CCCCAACCCTGGAAACAGAC-3′ and 5′-GGTCACGTTCAGTTGGTCAAAGG-3′ for mouse *Col1a1*; 5′-GGGCGTCTCCACAGTAACCG-3′ and 5′-ACTCCCACTGTGCCCTCGTT-3′ for mouse *Alpl*; 5′-CCAAGCAGGAGGGCAATA-3′ and 5′-AGGGCAGCACAGGTCCTAA-3′ for mouse *Bglap*; and 5′-ATGTGTCCGTCGTGGATCTG-3′ and 5′-TGAAGTCGCAGGAGACAACC-3′ for mouse *Gapdh*. Activation of osteoblasts was also evaluated by measuring the release of bone-formation markers into the medium using an ELISA kit for procollagen type I N-terminal propeptide (P1NP; Uscn Life Science) and an EIA kit for Gla-osteocalcin (TaKaRa) according to the manufacturer's instructions. To evaluate matrix mineralization by osteoblasts, cultured cells were fixed in 10% paraformaldehyde for 30 min at room temperature and stained using Alizarin Red S (Sigma-Aldrich; pH 4.2) or the von Kossa Method for Calcium Kit (Polysciences) as previously described[26]. Digital images were acquired with an optical scanner (GT-X970; EPSON).

Construction of viral vectors. The adenoviral vectors for transducing Runx2 or DN-Runx2 were constructed using pAd/CMV/V5-DEST vector (Life Technologies) and ViraPower Adenoviral Expression System (Life Technologies). DN-Runx2 consisted of the runt homology region with a nuclear localization sequence[27]. RANK-ORF or CD40-chimaera construct, a fusion protein of CD40 (Val24–Arg193) and RANK (Lys215–Glu615), were also subcloned into pAd/CMV/V5-DEST. β-galactosidase (LacZ) was used as reference adenoviral construct. To suppress endogenous RANK expression, an adenovirus carrying shRNA against the 3′-UTR of mouse RANK mRNA was constructed using a pAd/PL-DEST vector (Life Technologies). The target sequence was as follows: top strand: 5′-CACCGGGATGGCTTATTTAGCATACCGAAGTATGCTAAATAAGCCATCCC-3′ and bottom strand: 5′-AAAAGGGATGGCTTATTTAGCATACTTCGGTATGCTAAATAAGCCATCCC-3′.

Preparation of mOC-SEVs under conditions of RANK suppression. RAW 264.7 cells were seeded and stimulated with GST–sRANKL and M-CSF, as described above, for 24 h. Culture medium was replenished, and shRNA suppressing endogenous RANK was co-introduced with the CD40 chimaera or RANK-ORF using adenoviruses. Cells were additionally cultured for 96 h in the presence of 10 ng/ml M-CSF and 100 ng/ml mouse CD40 antibody (GeneTex) in the case of CD40-chimaera, or in the presence of 10 ng/ml M-CSF and 100 ng/ml GST–sRANKL in the case of RANK-ORF. To confirm osteoclast formation, tartrate-resistant acid phosphatase (TRAP) staining was performed using a commercial staining kit (Sigma-Aldrich). SEVs were collected by serial centrifugations and subjected to further analyses.

Immunoblot analyses of intracellular signalling proteins. For immunodetection of intracellular signalling molecules, cells were treated with stimulators, then lysed (in PBS, pH 7.4; 1.0% Triton X-100; protease inhibitors; and phosphatase inhibitors (Sigma-Aldrich)), and centrifuged at 15,000*g* for 10 min. The protein concentration of each supernatant was measured using a BCA Protein Assay Kit. Supernatants were used for immunoblot analyses. Extraction of nuclear proteins was conducted using NE-PER Nuclear and Cytoplasmic Extraction Reagents Kit (Thermo Scientific) according to the manufacturer's instructions. Blots were probed with the following primary antibodies: rabbit anti-phospho-Akt Thr308, rabbit anti-Akt, rabbit anti-phospho-PRAS40 Thr246, rabbit anti-PRAS40, rabbit anti-phospho-S6K1 Thr389, rabbit anti-S6K1, rabbit anti-GAPDH, rabbit anti-Hdac2 (Cell Signaling Technology), and mouse anti-Runx2 (MBL). Immunolabelled proteins were detected using HRP-labelled donkey anti-rabbit IgG or donkey anti-mouse IgG (GE Healthcare) as the corresponding secondary antibodies and ECL prime detection reagents (GE Healthcare).

Evaluation of osteogenesis in vivo. The osteogenic potency of mOC-SEV was evaluated using a mouse calvarial defect model according to previously described methods[28]. In brief, recombinant human BMP-2 (1 μg, Bioventus) or mOC-SEVs (50 μg) were added to gelatin hydrogel sheets (4 mm). A calvarial defect was generated on the left parietal bone of 5-week-old male C57BL/6 mice (Clea Japan) using a biopsy punch (Kai Industries; 3.5 mm); a gelatin hydrogel sheet was placed on the defect. To measure bone formation parameters, calcein (Sigma-Aldrich; 40 mg/kg) and demeclocycline (Sigma-Aldrich; 20 mg/kg) were administered on days 18 and 25 after implantation, respectively, and mice were killed on day 28. Calvariae were fixed in phosphate-buffered glutaraldehyde (0.25%)-formaldehyde (4%) fixative (pH 7.4) for 3 days at 4 °C. Soft X-ray images of calvariae were acquired using a cabinet X-ray apparatus (SRO-M50, Sofron Co. Ltd, Tokyo, Japan). Calvarial bone mineral density (BMD) and bone mineral content (BMC) were measured by dual-energy X-ray absorptiometry (DXA, DCS-600R, ALOKA). Standard bone histomorphometric analyses were performed using an image analysing system (KS400, Carl Zeiss) as described previously[29].

Screening of scFv targeting for RANKL and IGSF8. Clones of scFv recognizing RANKL ectodomain were isolated from CDR2/3-randomized human-scFv phage library, Tomlinson I + J (DNAFORM), following the manufacturer's protocol with minor modifications[30]. Phage at 1×10^{12} plaque forming units (pfu) were used for each round of panning. After pre-adsorption treatment with GST (immobilized on glutathione sepharose 4B beads), phage were incubated with immobilized

GST–sRANKL for 2 h at 4 °C. Beads were washed with washing buffer and PBS, three times, respectively. Phage were eluted with Tris-EDTA buffer (pH 8.0) containing 20 mM glutathione. The obtained phage were added to TG1 *E. coli* in the exponential growth phase. Transformed TG1 cells were cultured in 2× YT medium, and amplified phage particles were precipitated using polyethylene glycol after infection with helper phage. After three cycles of panning, phage clones were isolated. To obtain scFv clones recognizing the IGSF8 ectodomain (Met1–Tyr577), fusion protein with human IgG1 Fc region was used as bait, and phage display screening was performed as described above. As a reference clone, anti-BSA scFv was used as recommended. The binding specificity of each clone to GST–sRANKL or IGSF8-Fc was confirmed using ELISA as follows. In brief, multiwell plates were coated with GST–sRANKL or IGSF8-Fc (2 μg/ml in 50 mM sodium carbonate buffer, pH 9.6) for 12 h at 4 °C, and 2% skim milk in PBS was used in the subsequent blocking step for 2 h at room temperature. After three washes, diluted phage specimens in 2% skim milk were incubated for 2 h at room temperature. After three washes, HRP-labelled mouse anti-phage M13 antibody (GE Healthcare; diluted 1:5,000 in 2% skim milk) was incubated for 1 h at room temperature. After five washes, colorimetric detection was performed as described above.

Preparation and characterization of scFv trimers. To obtain recombinant scFv in a trimeric form, an isoleucine zipper (MKQIEDKIEEILSKIYHIENE IARIKKLIGER[31]) was fused to the C terminus of scFv with a 6×His tag connected by a short linker sequence (GSTSG). In addition, a signal peptide (MKYLLPTAAAGLLLLAAQPA) was fused at the N terminus to ensure efficient protein translocation to the periplasm. This construct was subcloned into an *E. coli* expression vector, pQE-2 (QIAGEN, Hilden, Germany), and transformed into Rosetta 2 *E. coli* cells. Transformed cells were cultured in 2× YT medium at 37 °C until the $A_{600\ \mathrm{nm}}$ reached 0.8. IPTG was added at a final concentration of 125 μM and additionally cultured for 24 h at 25 °C. Cells were collected and suspended in hypertonic buffer-A (100 mM Tris-HCl buffer, pH 8.0, 20% sucrose, 5 mM EDTA) on ice for 30 min. After centrifugation, cell pellets were resuspended in hypotonic buffer-B (5 mM $\mathrm{MgCl_2}$) for 10 min and centrifuged again. Supernatants of buffer-A and buffer-B were collected, and PD-10 columns were used to exchange buffer constituents (50 mM phosphate buffer, pH 8.0, 300 mM NaCl, 15 mM imidazole) in order to perform purification with Profinity Ni-IMAC (Bio-Rad). The 6×His-tagged scFv trimer protein was purified according to the manufacturer's protocol and preserved in saline. Protein concentration and purity were confirmed by SDS–PAGE followed by staining with Coomassie blue. We confirmed that endotoxin contamination was negligible using a conventional limulus amebocyte lysate test kit. The binding affinity of the scFv trimers to RANKL or IGSF8 was determined by ELISA using a rabbit anti-6×His-tag antibody (Rockland Immunochemicals; 33 ng/ml) and HRP-labelled donkey anti-rabbit IgG (GE Healthcare; 1:5,000 dilution).

RANKL(Pro29Ala) mice. RANKL(Pro29Ala) mutant mice were generated by Unitech. In brief, a targeting vector was constructed using a genomic fragment containing exon 1 of the mouse *RANKL* gene (also known as *Tnfsf11*). A Pro29(CCG) to Ala(gCG) point mutation was introduced by site-directed mutagenesis using a QuikChange kit (Agilent Technologies); a loxP-flanked neomycin-resistance cassette was inserted into a downstream intron. This targeting vector was introduced into C57BL/6-derived embryonic stem (ES) cells, and clones positive for homologous recombination at the RANKL locus were selected. Blastocysts were injected with ES cells and implanted into pseudo-pregnant female mice. Germline transmitted chimeric mice were subsequently bred with CAG-Cre transgenic mice to eliminate the neomycin-resistance cassette. An additional cross with C57BL/6 wild-type mice was performed to eliminate the Cre transgene and obtain mice carrying the $RANKL^{P29A}$ mutation and the inserted loxP nucleotides in the downstream intron.

Evaluation of the coupling process in vivo. To evaluate the influence of RANKL reverse signalling on the coupling of bone resorption and formation, 8-week-old female $RANKL^{P29A}$ homozygous mutant or wild-type littermates were administered GST–sRANKL (1 mg/kg) intraperitoneally three times every 24 h (days 1 to 3) and their femurs were analysed on day 13. The influence of the $RANKL^{P29A}$ mutation on physiological bone phenotypes was examined using $RANKL^{P29A};OPG^{+/-}$ mice and $RANKL^{WT};OPG^{+/-}$ littermates. Lumbar spines in 7-week-old female mice were analysed. To evaluate the influence of suppressing mOC-SEVs, 8-week-old female wild-type mice were administered GST–sRANKL as described above, and GW4869 (Sigma-Aldrich; 2.5 mg/kg) or αI-tri1/2 (10 mg/kg) was administered every 12 h for 5 days (days 0 to 4). In all experiments, demeclocycline (20 mg/kg) and calcein (40 mg/kg) were injected subcutaneously 7 and 2 days before sacrifice, respectively. Femurs or lumbar spines were fixed in PBS containing 4% formaldehyde for 3 days at 4 °C. Standard bone histomorphometric analyses were performed using an image analysis system (KS400) as described previously[29].

Construction of αR-bif and αR-ref. To sufficiently multimerize RANKL, heavy- to light-chain oriented scFv against RANKL, connected with a shorter linker sequence (GGG), was tandemly repeated via α-helix linker (DWSNSEEAKKEEAKKEEAKKSNSV). Human IgG1 Fc-region was fused to the C terminus of this scFv diabody to improve the pharmacokinetic property (αR-bif). As a reference protein, light- to heavy-chain oriented scFv against RANKL, connected with a longer linker sequence (GGGGSGGGGSGGGGS), was also tandemly repeated via α-helix linker, and human IgG1 Fc-region was fused to the C terminus of this tandem scFv (αR-ref). Both chimeric constructs were subcloned into pcDNA3.3 vector (Life Technologies), and transduced into 293FT cells using PEI MAX (Polysciences) according to the manufacturer's protocol. Culture medium was changed to serum-free medium (OPTIPRO SFM, Life Technologies), containing 10 mM sodium butyrate, 24 h after the transfection and additionally cultured for 120 h. Cell debris was removed by centrifugation. Then, the supernatant was collected, and Fc-fusion proteins were purified using Protein A sepharose beads (GE Healthcare) according to the manufacturer's protocol. After buffer exchange of the eluted sample to saline using PD-10 columns, the protein concentration was measured using a BCA Protein Assay Kit.

Characterization of αR-bif and αR-ref in vitro. The binding affinity of αR-bif or αR-ref to the RANKL ectodomain was determined by ELISA. In brief, multiwell plates were coated with goat anti-human IgG Fc antibody (Sigma-Aldrich; 2 μg/ml in 50 mM sodium carbonate buffer, pH 9.6) for 12 h at 4 °C, and 5% skimmed milk was used in the blocking step for 2 h at room temperature. After three washes, αR-bif or αR-ref (400 nM) was added and incubated for 12 h at 4 °C. After three washes, diluted recombinant soluble RANKL with 6×His tag was added and incubated for 1 h at 37 °C. After five washes, HRP-labelled mouse anti-6×His-tag antibody (GE Healthcare; 1:5,000 dilution) was used for the detection. To evaluate the inhibitory effect on osteoclastogenesis, RAW 264.7 cells were seeded in 96-well plates and cultured with 100 ng/ml GST–sRANKL and 10 ng/mL M-CSF for 96 h. αR-bif or αR-ref was simultaneously added to the culture medium, and TRAP enzymatic activity was determined using a TRACP & ALP Assay Kit (TaKaRa) following the manufacturer's protocol. Agonistic action on RANKL reverse signalling was evaluated using ST2 cells as described above.

Evaluation of the pharmacological effect in OVX mice. To induce an osteoporotic condition in vivo, ovariectomy was conducted in 7-week-old female C57BL/6 mice as reported previously[32]. Purified αR-bif or αR-ref was administered intravenously on day 16 after the surgery. The fluorescent labelling reagents demeclocycline, calcein, and alizarin complexone (Sigma-Aldrich; 30 mg/kg) were injected subcutaneously on days 19, 24, and 29, respectively. Mice were killed on day 31, and standard bone histomorphometric analyses of femurs were performed.

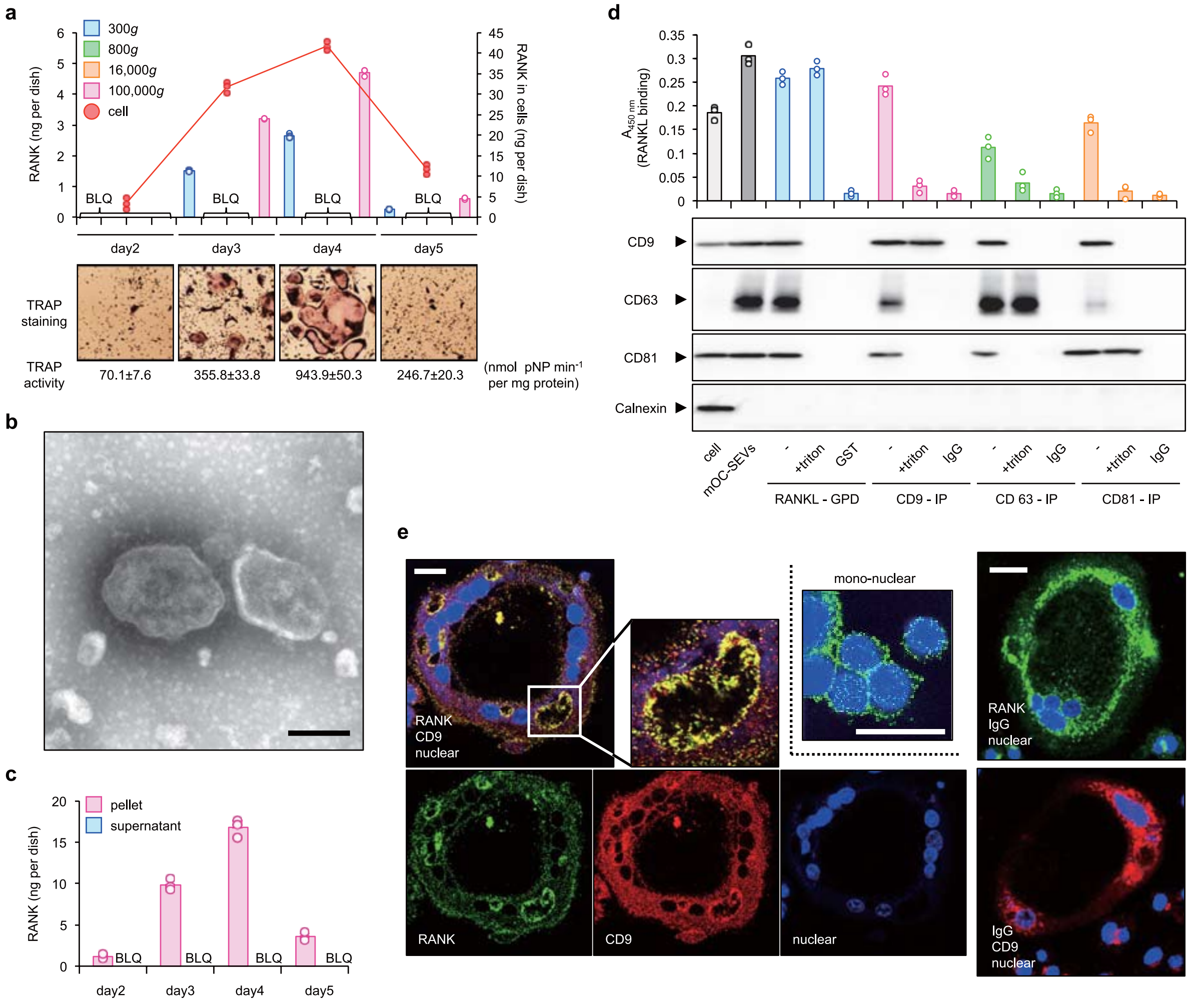
Statistical analysis and reproducibility. No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment. All quantitative data are expressed as the mean, represented as column bar or line graphs, with individual biologically independent values as scatter plots. Statistical analysis was performed using unpaired two-sided Student's *t*-test, or one-way analysis of variance followed by the Student–Newman–Keuls test to compare three or more groups. Differences were considered significant when $P < 0.05$; $^{*}P < 0.05$, $^{**}P < 0.01$. Images of western blot analyses, von-Kossa staining and others show one experiment which is representative of repeated independent experiments, and reproducibility was validated as follows. Reproducibility of results in Fig. 2a, Extended Data Figs. 2a, e, f, 3a, c, e, 5d were confirmed by two independent experiments, and in Figs. 1e, f, 2c, d, f, g, 3c, f, 4c, 5b, Extended Data Figs. 1a, b, d, e, 2d, 4a–c, 5a, b, 6a, b, e, f, 7g, 8a–d, f, g were validated by at least three independent experiments. Information about statistical analysis is mentioned in each figure legend.

Animal procedures. All animal experiments were performed in accordance with the relevant ethical guidelines and regulations. All procedures were approved by the Institutional Animal Care and Use Committee of the Graduate School of Medicine, the University of Tokyo, the Animal Care and Use Committee of Tokyo Medical and Dental University, or the Animal Management Committee of Matsumoto Dental University.

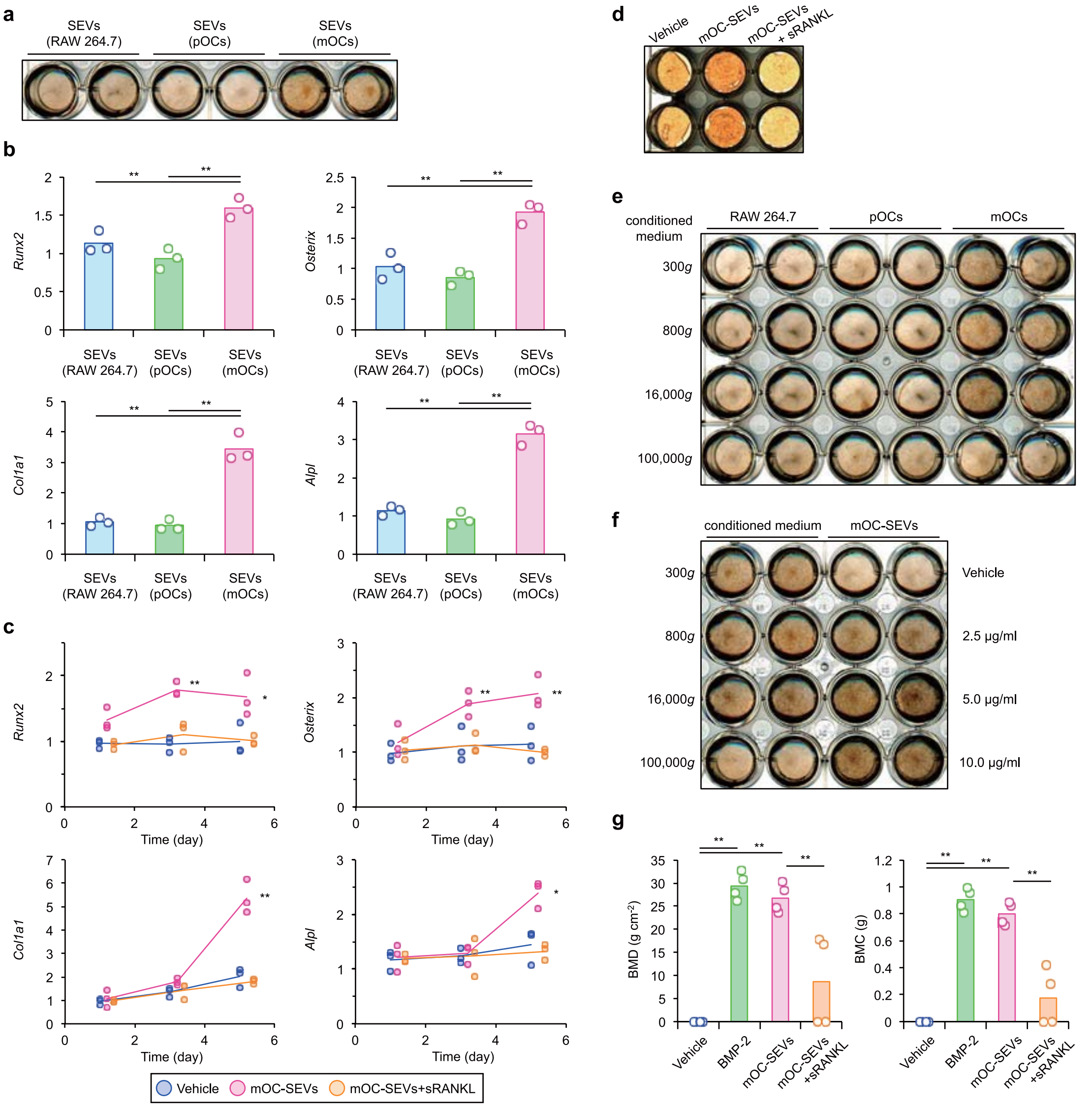
Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability. Source data for the figures and extended data figures are provided in the online version of the paper. All relevant data regarding extracellular vesicles have been deposited to the EV-TRACK knowledge base (http://evtrack.org)[33] under accession number EV180042.The mass spectrometry proteomics data analysing the content of mOC-SEVs, related to the list of identified proteins shown in Supplementary Table 1, have been deposited to the ProteomeXchange Consortium via the PRIDE[34] partner repository with the dataset identifier PXD010500. Uncropped images of western blots and SDS–PAGE analyses are available in the online version of this paper as Supplementary Fig. 1. Detailed description and methods for Extended Data Figures are provided in Supplementary Information (see also Supplementary Information Guide).

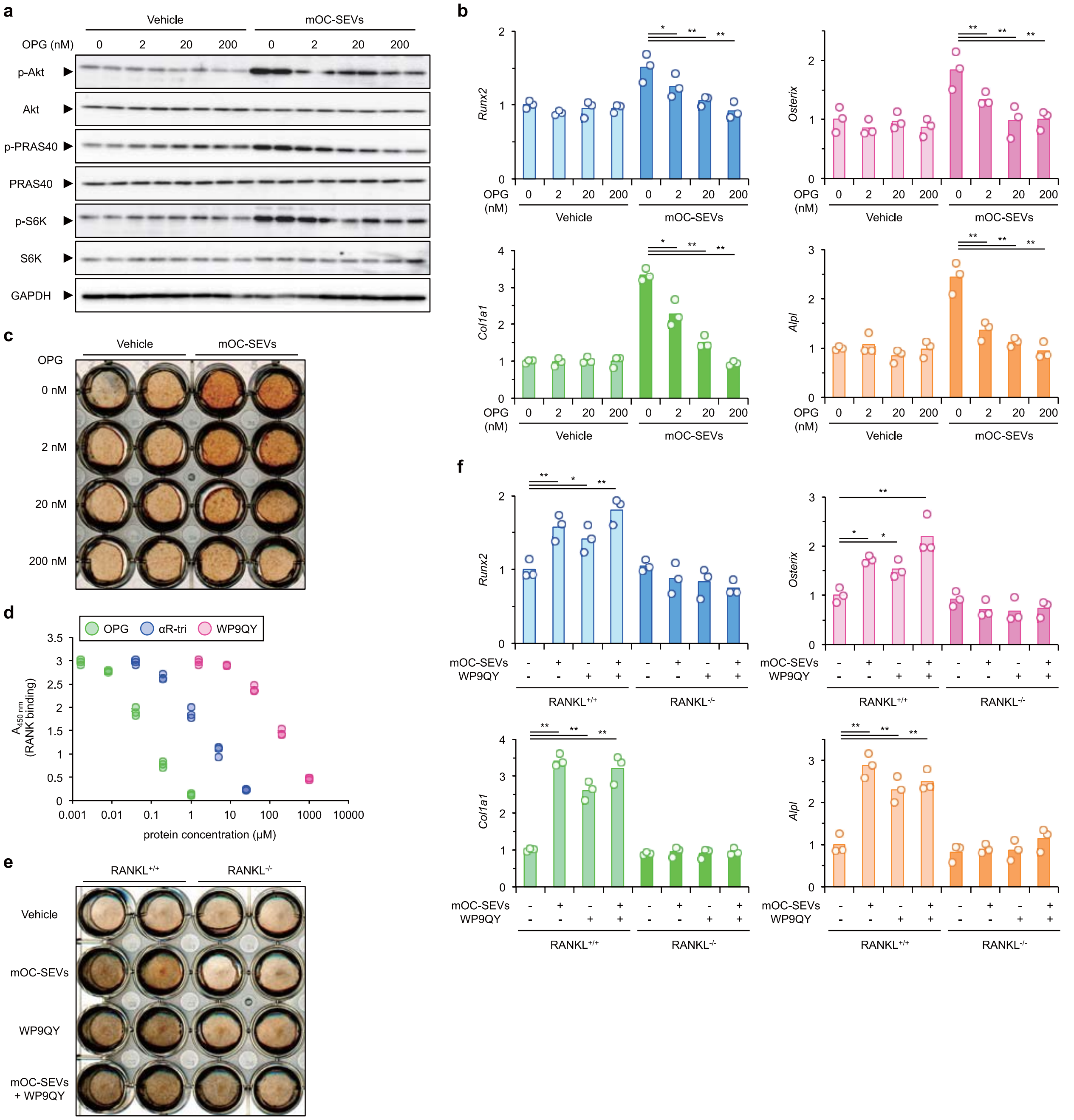
25. Sato, N. et al. MyD88 but not TRIF is essential for osteoclastogenesis induced by lipopolysaccharide, diacyl lipopeptide, and IL-1α. *J. Exp. Med.* **200**, 601–611 (2004).
26. George, J., Kuboki, Y. & Miyata, T. Differentiation of mesenchymal stem cells into osteoblasts on honeycomb collagen scaffolds. *Biotechnol. Bioeng.* **95**, 404–411 (2006).
27. Maruyama, Z. et al. Runx2 determines bone maturity and turnover rate in postnatal bone development and is involved in bone loss in estrogen deficiency. *Dev. Dyn.* **236**, 1876–1890 (2007).
28. Alles, N. et al. Suppression of NF-κB increases bone formation and ameliorates osteopenia in ovariectomized mice. *Endocrinology* **151**, 4626–4634 (2010).
29. Dempster, D. W. et al. Standardized nomenclature, symbols, and units for bone histomorphometry: a 2012 update of the report of the ASBMR Histomorphometry Nomenclature Committee. *J. Bone Miner. Res.* **28**, 2–17 (2013).
30. Pini, A. et al. Design and use of a phage display library. Human antibodies with subnanomolar affinity against a marker of angiogenesis eluted from a two-dimensional gel. *J. Biol. Chem.* **273**, 21769–21776 (1998).
31. Watzl, C. Production and use of trimeric isoleucine zipper fusion proteins to study surface receptor ligand interactions. *Curr. Protoc. Protein Sci.* **19**, 19.11.1–19.11.13 (2006).
32. Sakai, A. et al. Bone marrow cell development and trabecular bone dynamics after ovariectomy in ddy mice. *Bone* **23**, 443–451 (1998).
33. Van Deun, J. et al. EV-TRACK: transparent reporting and centralizing knowledge in extracellular vesicle research. *Nat. Methods* **14**, 228–232 (2017).
34. Vizcaino, J. A. et al. 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res.* **44**, D447–D456 (2016).



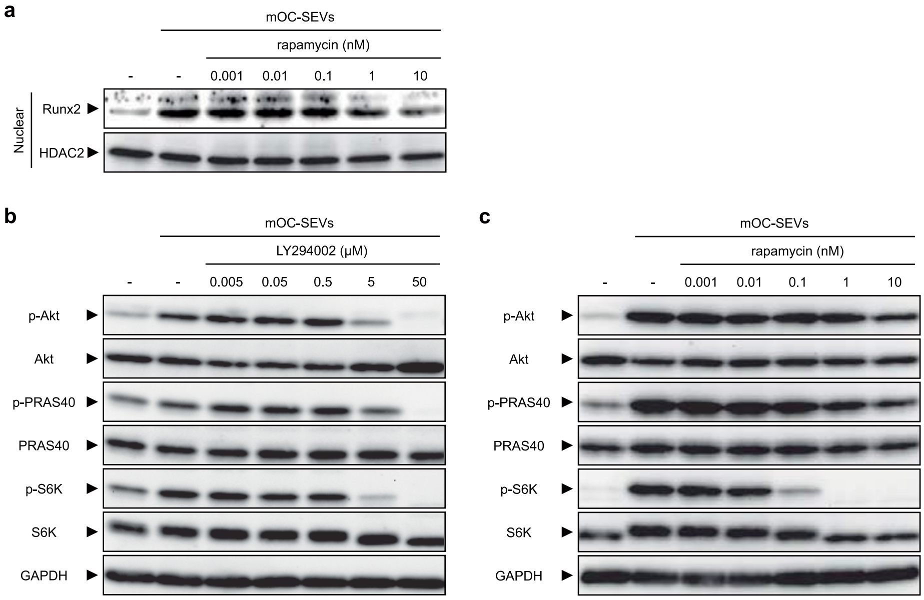
Extended Data Fig. 1 | Incorporation of RANK into mOC-SEVs. **a**, RANK content over time (top) in osteoclasts (lines) and their secretion products that were fractionated by serial centrifugation (columns) (*n* = 3); BLQ, below the limit of quantification. Bottom, osteoclast maturation over time with images of TRAP-stained cells and TRAP activity (*n* = 4). **b**, Transmission electron micrograph of precipitates from culture supernatants of mOCs by ultracentrifugation at 100,000*g*. Scale bar, 100 nm. **c**, RANK content in SEVs collected from osteoclast culture supernatants over time using ExoQuick-TC SEV-precipitating reagent (*n* = 3). **d**, Co-precipitation analyses of mOC-SEVs. Top, RANK content in co-precipitates (*n* = 3). Bottom, co-precipitates analysed for SEV markers (CD9, CD63, and CD81) and an endoplasmic reticulum protein (calnexin) by immunoblotting. GST pull-down (GPD) assays were used to precipitate RANK, and immuno-precipitation (IP) was used to precipitate CD9, CD63, and CD81, respectively. **e**, Confocal laser scanning micrograph of osteoclasts labelled for RANK (green), CD9 (red), and nuclei (blue). Scale bar, 20 μm. In **a**, **c**, **d**, data are mean, with individual points representing biologically independent samples.



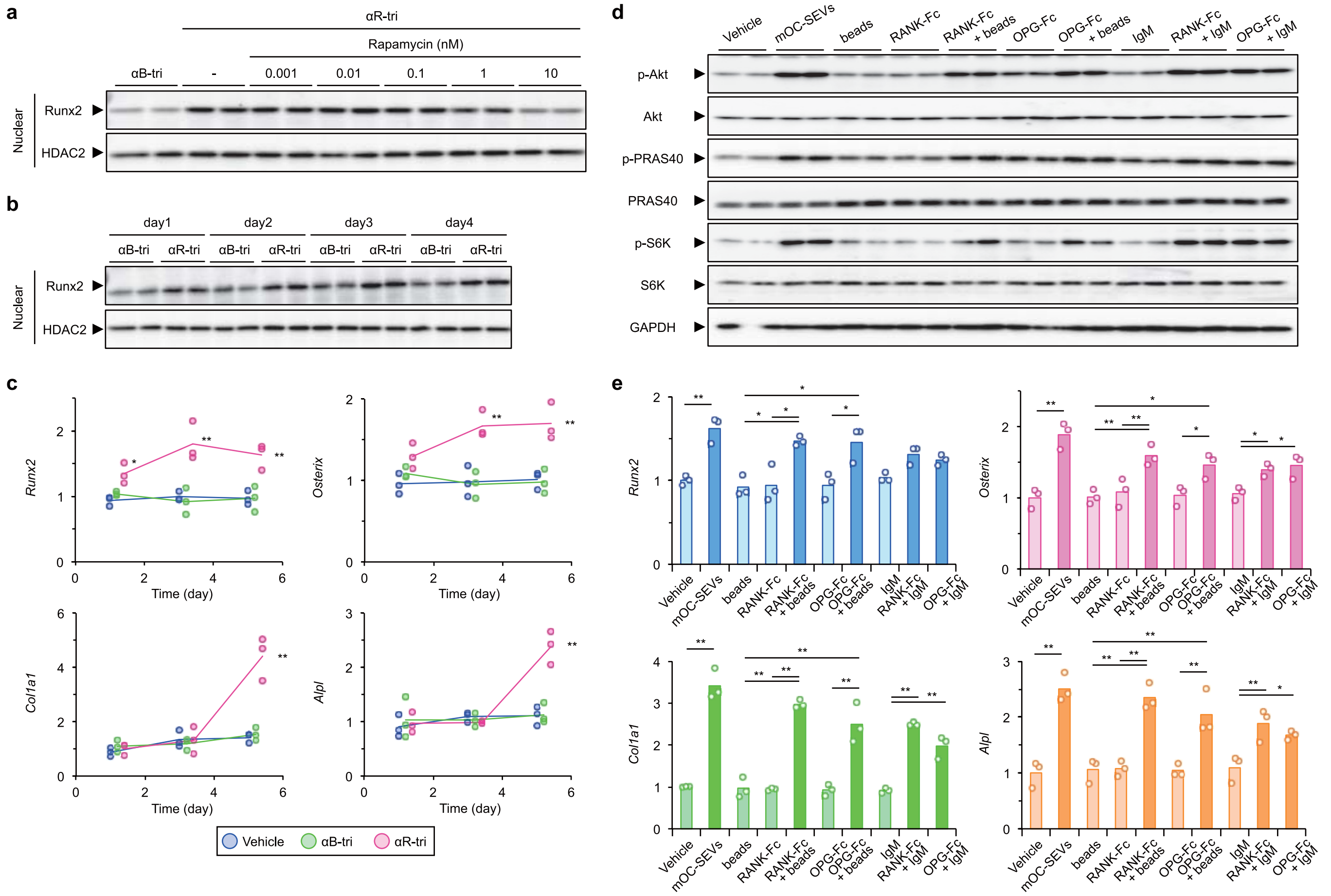
Extended Data Fig. 2 | Osteogenic potency of mOC-SEVs under physiological conditions. a, Images of von Kossa-stained ST2 cells stimulated with SEVs derived from mOCs, pOCs, or unstimulated RAW 264.7 cells. **b**, Changes in mRNA expression levels of osteoblast markers in ST2 cells after 5 days of stimulation with SEVs derived from mOCs, pOCs, or unstimulated RAW 264.7 cells ($n = 3$). **c**, Expression of osteoblast markers in mouse primary osteoblasts stimulated with mOC-SEVs, mOC-SEVs plus sRANKL, or vehicle ($n = 3$). **d**, Von Kossa-stained primary osteoblasts stimulated with mOC-SEVs, mOC-SEVs plus sRANKL, or vehicle. **e**, Von Kossa-stained ST2 cells stimulated with serially-centrifuged conditioned medium from mOCs, pOCs, or unstimulated RAW 264.7 cells. **f**, Comparison of osteoblast-stimulating effect of conditioned medium from mOCs with that of purified mOC-SEVs. Von Kossa-stained ST2 cells stimulated with serially-centrifuged conditioned medium from mOCs, or with 2.5, 5.0, and 10.0 µg protein per ml purified mOC-SEVs. **g**, Dual-energy X-ray absorptiometry measurements of bone mineral density (BMD) and bone mineral content (BMC). Measurements were performed on the injured skull region of the mouse calvarial defect model after treatment with mOC-SEVs, mOC-SEVs plus sRANKL, BMP-2 (positive control), or vehicle ($n = 4$). In **b**, **c**, **g**, data are mean, with individual scatter points representing biologically independent samples or mice. One-way ANOVA followed by Student–Newman–Keuls test; $*P < 0.05$, $**P < 0.01$.



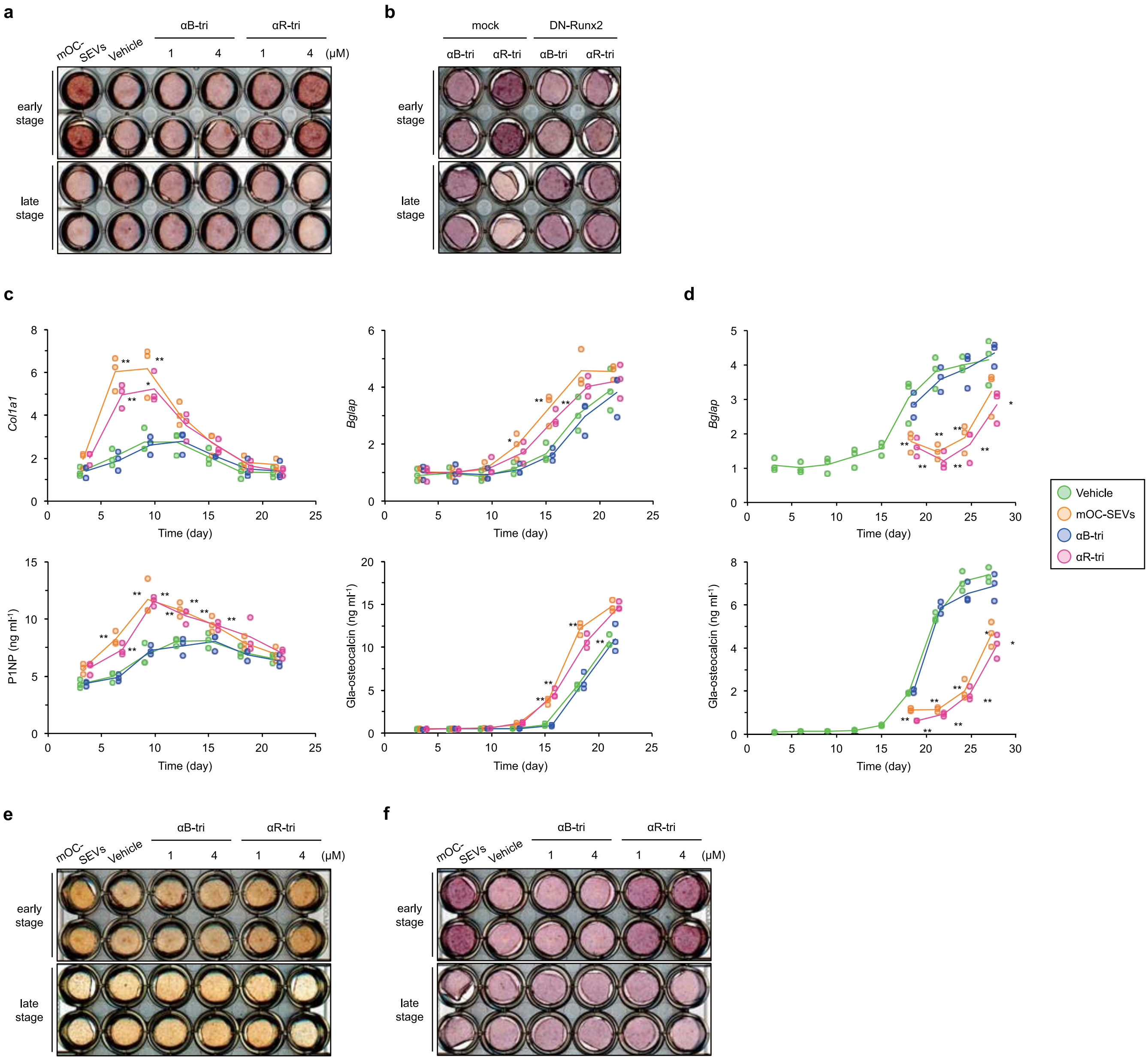
Extended Data Fig. 3 | Influence of RANKL-binding molecules on the osteogenic effects of mOC-SEVs. **a**, The inhibitory effect of OPG on the activation of PI3K–Akt–mTORC1 pathway in ST2 cells stimulated with mOC-SEVs. Akt Thr308 phosphorylation (p-Akt), PRAS40 Thr246 phosphorylation (p-PRAS40) and S6K1 Thr389 phosphorylation (p-S6K1) indicate PI3K, Akt and mTORC1 activation, respectively. **b**, The inhibitory effect of OPG on osteoblast-marker expression levels in ST2 cells after 5 days of stimulation with mOC-SEVs ($n = 3$). **c**, Von Kossa-stained ST2 cells stimulated with mOC-SEVs or vehicle in the presence or absence of OPG. **d**, Inhibitory effect of the WP9QY peptide, αR-tri, or OPG on RANK-Fc binding to GST–sRANKL immobilized on ELISA plate ($n = 3$ biologically independent samples). **e**, Von Kossa-stained primary osteoblasts derived from RANKL-deficient mice or wild-type littermates stimulated with mOC-SEVs, WP9QY peptide, mOC-SEVs plus WP9QY peptide, or vehicle. **f**, mRNA expression levels of osteoblast markers in primary osteoblasts derived from RANKL-deficient mice or wild-type littermates after 5 days of stimulation with mOC-SEVs, WP9QY peptide, mOC-SEVs plus WP9QY peptide, or vehicle ($n = 3$). In **b**, **f**, data are mean, with individual points representing biologically independent samples. One-way ANOVA followed by Student–Newman–Keuls test; $*P < 0.05$, $**P < 0.01$.


a
mOC-SEVs
rapamycin (nM)
-
-
0.001
0.01
0.1
1
10
Nuclear
Runx2
HDAC2
b
mOC-SEVs
LY294002 (μM)
-
-
0.005
0.05
0.5
5
50
p-Akt
Akt
p-PRAS40
PRAS40
p-S6K
S6K
GAPDH
c
mOC-SEVs
rapamycin (nM)
-
-
0.001
0.01
0.1
1
10
p-Akt
Akt
p-PRAS40
PRAS40
p-S6K
S6K
GAPDH

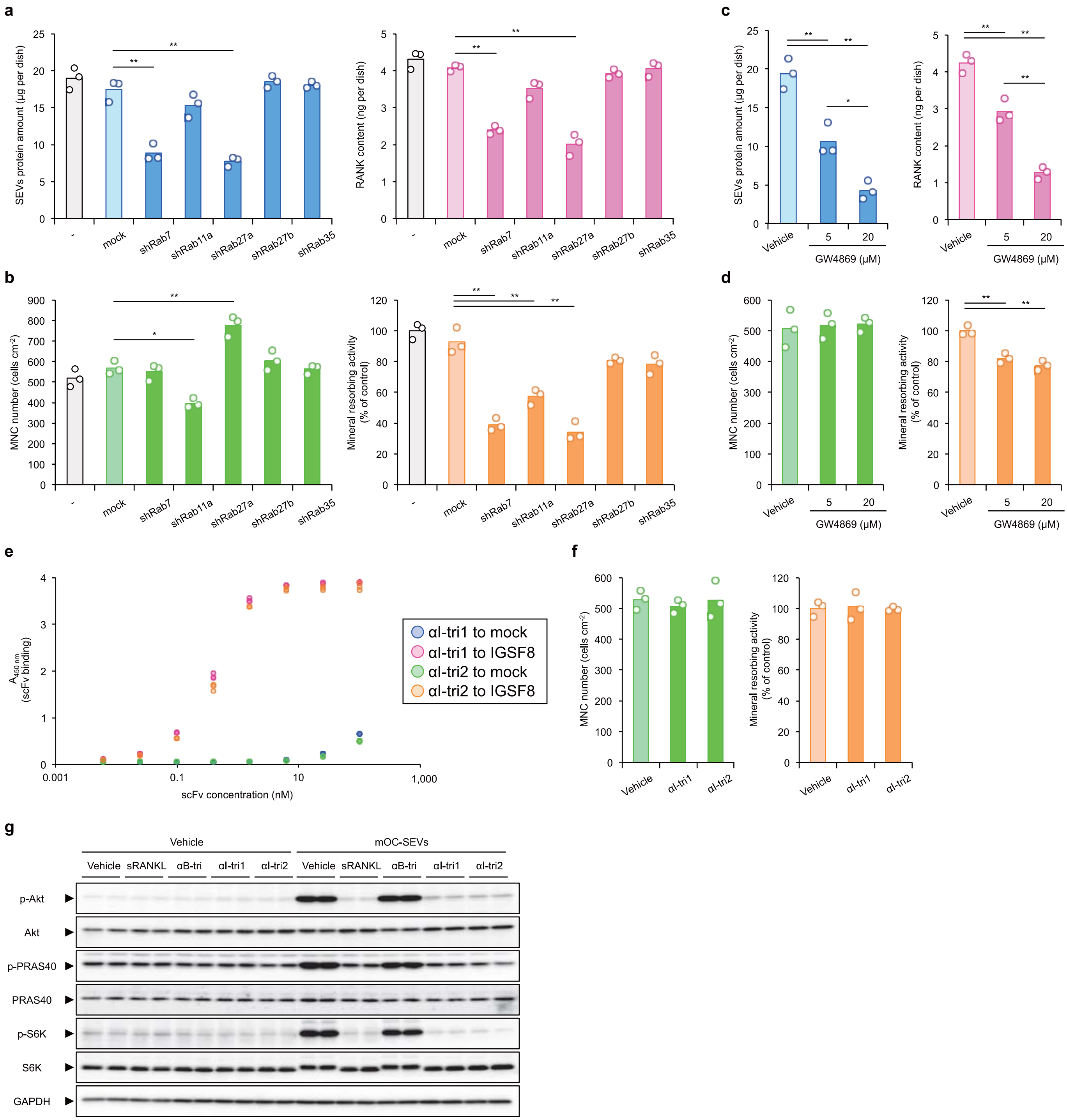
Extended Data Fig. 4 | Activation of PI3K–Akt–mTORC1–Runx2 axis by vesicular RANK. a, The effect of rapamycin on nuclear translocation of Runx2 in ST2 cells stimulated with mOC-SEVs. Histone deacetylase 2 (Hdac2) was used as reference nuclear protein. **b**, **c**, Influence of PI3K inhibitor (LY294002) (**b**) or mTORC1 inhibitor (rapamycin) (**c**) on activation of the PI3K–Akt–mTORC1 pathway in ST2 cells stimulated with mOC-SEVs.



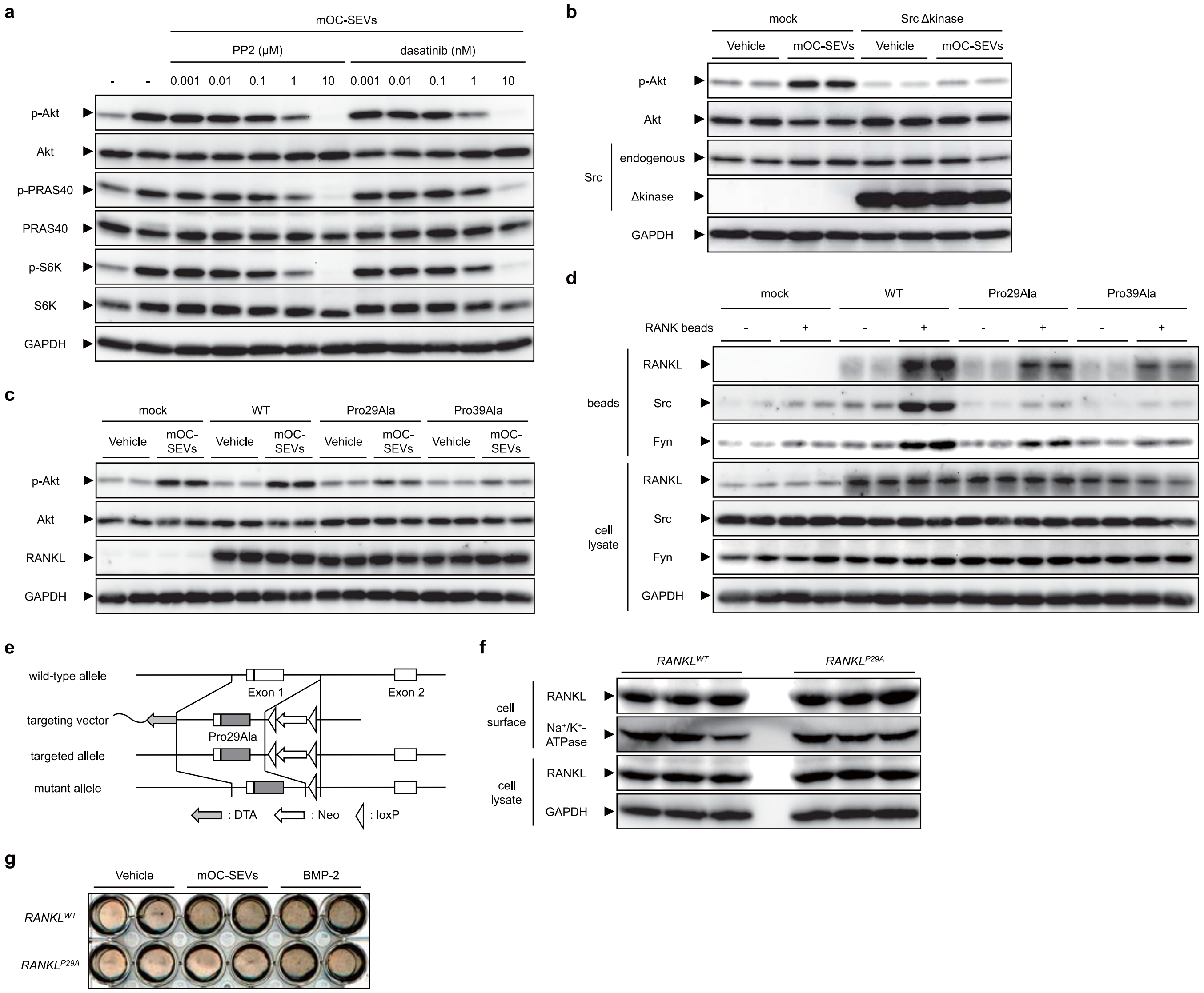
Extended Data Fig. 5 | Osteoblast-stimulating effects of RANKL-crosslinking agents. a, Influence of mTORC1 inhibitor (rapamycin) on nuclear Runx2 in ST2 cells stimulated with αR-tri or αB-tri. **b**, Time course of nuclear Runx2 in mouse primary osteoblasts stimulated with αR-tri or αB-tri. **c**, Time course of osteoblastic marker expression in mouse primary osteoblasts stimulated with αR-tri, αB-tri, or vehicle ($n = 3$). **d**, Activation of PI3K–Akt–mTORC1 pathway in ST2 cells stimulated with RANKL-crosslinking agents, RANK-Fc plus beads, OPG-Fc plus beads, RANK-Fc plus IgM, and OPG-Fc plus IgM. Vehicle, RANK-Fc, OPG-Fc, beads, and IgM were used as respective controls. 10 μg ml^{-1} mOC-SEVs (containing about 1 ng ml^{-1} RANK) and 1 μg ml^{-1} RANK-Fc or OPG-Fc was used. **e**. Expression of osteoblast markers in ST2 cells after 5 days of stimulation with RANKL-crosslinking agents. ($n = 3$). **c**, **e**, Data are mean, with individual points representing biologically independent samples. One-way ANOVA followed by Student–Newman–Keuls test; $*P < 0.05$, $**P < 0.01$.



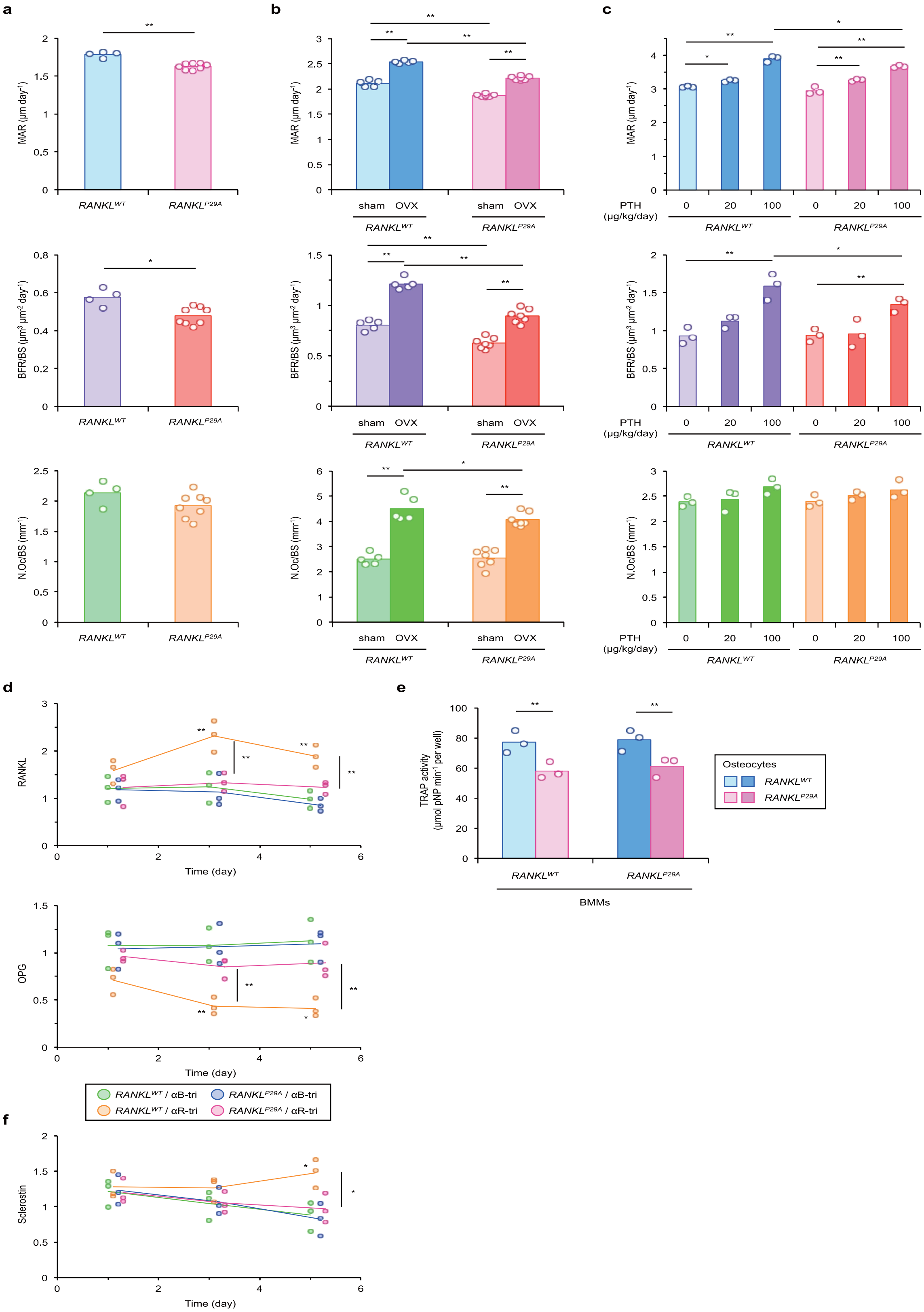
Extended Data Fig. 6 | Confirmation of biphasic property of RANKL reverse signalling. **a**, Alizarin red staining of ST2 cells stimulated with mOC-SEVs, αR-tri, αB-tri, or vehicle during the early or late stage of differentiation. **b**, Alizarin red-stained ST2 cells infected with DN-Runx2 adenovirus before stimulation with αR-tri or αB-tri during the early or late stage of differentiation. **c**, **d**, Time courses of expression of early osteoblastic markers (*Col1a1* mRNA and P1NP value reflecting the expression of type I collagen protein) and late osteoblastic markers (osteocalcin mRNA and Gla-osteocalcin protein) in mouse primary osteoblasts stimulated during the early stage (**c**) or late stage (**d**) of differentiation with mOC-SEVs, αR-tri, αB-tri, or vehicle ($n = 3$). **e**, **f**, Von Kossa-stained (**e**) or alizarin red-stained (**f**) mouse primary osteoblasts stimulated with mOC-SEVs, αR-tri, αB-tri, or vehicle during the early stage or late stage of differentiation. In **c** and **d**, data are mean, with individual points representing biologically independent samples.



Extended Data Fig. 7 | The effect of neutralizing mOC-SEVs. **a**, Influence of suppressing mRNA expression of Rab family members in osteoclasts on the secretion of vesicular RANK from mOCs ($n = 3$). **b**, Influence of suppressing mRNA expression of Rab family members in osteoclasts on formation of multinuclear cells (MNCs) and their mineral-resorbing activity ($n = 3$). **c**, Influence of GW4869 on the secretion of vesicular RANK from mOCs ($n = 3$). **d**, Influence of GW4869 on formation of MNCs and their mineral-resorbing activity ($n = 3$). **e**, Binding affinity of αI-tri1 or αI-tri2 to the IGSF8 ectodomain ($n = 3$ biologically independent samples). **f**, Influence of αI-tri1/2 on formation of MNCs and their mineral-resorbing activity ($n = 3$). **g**, Influence of pre-incubating mOC-SEVs with αI-tri1/2, αB-tri, GST–sRANKL or vehicle on the stimulatory activity of mOC-SEVs on ST2 cells; activation of the PI3K–Akt–mTORC1 pathway in ST2 cells is shown. In **a**–**d**, **f**, data are mean, with individual points representing biologically independent samples. Statistical analyses, one-way ANOVA followed by Student–Newman–Keuls test; $^{*}P < 0.05$, $^{**}P < 0.01$.


a
mOC-SEVs
PP2 (μM)
dasatinib (nM)
- - 0.001 0.01 0.1 1 10 0.001 0.01 0.1 1 10
p-Akt
Akt
p-PRAS40
PRAS40
p-S6K
S6K
GAPDH
b
mock
Src Δkinase
Vehicle
mOC-SEVs
Vehicle
mOC-SEVs
p-Akt
Akt
Src
endogenous
Δkinase
GAPDH
c
mock
WT
Pro29Ala
Pro39Ala
Vehicle
mOC-SEVs
p-Akt
Akt
RANKL
GAPDH
d
mock
WT
Pro29Ala
Pro39Ala
RANK beads
- + - + - + - +
beads
RANKL
Src
Fyn
cell lysate
RANKL
Src
Fyn
GAPDH
e
wild-type allele
Exon 1
Exon 2
targeting vector
Pro29Ala
targeted allele
mutant allele
: DTA
: Neo
: loxP
f
RANKL^WT
RANKL^P29A
cell surface
RANKL
Na+/K+-ATPase
cell lysate
RANKL
GAPDH
g
Vehicle
mOC-SEVs
BMP-2
RANKL^WT
RANKL^P29A

Extended Data Fig. 8 | Effect of introducing RANKL Pro29Ala mutation. **a**, Src family kinase inhibitors (PP2 or dasatinib) inhibit the activation of the PI3K–Akt–mTORC1 pathway in ST2 cells stimulated with mOC-SEVs. **b**, Overexpression of kinase-dead Src mutant (Δkinase) inhibits PI3K activation in ST2 cells stimulated with mOC-SEVs. **c**, Overexpression of RANKL mutants (Pro29Ala or Pro39Ala) inhibits PI3K activation in ST2 cells stimulated with mOC-SEVs. **d**, Co-precipitation of Src and Fyn with RANKL in ST2 cells stimulated with RANK beads. RANKL Pro29Ala or Pro39Ala mutations decrease co-precipitation. **e**, Schematic diagram depicting the introduction of RANKL Pro29Ala point mutation. DTA, diphtheria toxin subunit A; *Neo*, neomycin-resistance gene. **f**, Expression and cell-surface presentation of RANKL in primary osteoblasts derived from wild-type and *RANKL*^{P29A} mutant mice. **g**, Von Kossa-stained primary osteoblasts derived from *RANKL*^{P29A} mutant mice and wild-type littermates stimulated with mOC-SEVs.



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Phenotypes of *RANKL*^{P29A} mutant mice. a, Bone morphometric parameters in 17-week-old male *RANKL*^{P29A} mice ($n = 8$) and wild-type littermates ($n = 4$) under physiological conditions. **b**, Bone morphometric parameters in OVX or sham operated 16-week-old female *RANKL*^{P29A} mice ($n = 7$ (sham), 7 (OVX)) and wild-type littermates ($n = 5$ (sham), 5 (OVX)). **c**, Bone morphometric parameters in 8-week-old male *RANKL*^{P29A} mice and wild-type littermates treated with daily subcutaneous injection of 20 or 100 μg kg^{-1} day^{-1} parathyroid hormone (PTH) over the course of 2 weeks ($n = 3$ in each group). **d**, Expression levels over time of RANKL and OPG mRNA in primary osteocytes derived from *RANKL*^{P29A} mice and wild-type littermates stimulated with αR-tri or αB-tri ($n = 3$). **e**, TRAP activity in bone marrow macrophages (BMMs) and primary osteocyte co-culture derived from *RANKL*^{P29A} mice and wild-type littermates ($n = 3$). **f**, Time-courses of the expression levels of sclerostin mRNA in primary osteocytes derived from *RANKL*^{P29A} mutant mice and wild-type littermates stimulated with αR-tri or αB-tri ($n = 3$). In **a**–**f**, data are mean, with individual points representing biologically independent samples or mice. Unpaired two-sided Student's *t*-test (**a**) or one-way ANOVA followed by Student–Newman–Keuls test (**b**–**f**.); $^{*}P < 0.05$, $^{**}P < 0.01$.

https://doi.org/10.1038/s41586-018-0481-8

Precancerous neoplastic cells can move through the pancreatic ductal system

Alvin P. Makohon-Moore[1,15], Karen Matsukuma[2,15], Ming Zhang[3,15], Johannes G. Reiter[4,5,15], Jeffrey M. Gerold[5,15], Yuchen Jiao[6], Lisa Sikkema[1,7], Marc A. Attiyeh[1], Shinichi Yachida[8], Corinne Sandone[9], Ralph H. Hruban[3,10,11], David S. Klimstra[12], Nickolas Papadopoulos[6], Martin A. Nowak[5,13], Kenneth W. Kinzler[6], Bert Vogelstein[3,6,14] & Christine A. Iacobuzio-Donahue[1,12]*

Most adult carcinomas develop from noninvasive precursor lesions, a progression that is supported by genetic analysis. However, the evolutionary and genetic relationships among co-existing lesions are unclear. Here we analysed the somatic variants of pancreatic cancers and precursor lesions sampled from distinct regions of the same pancreas. After inferring evolutionary relationships, we found that the ancestral cell had initiated and clonally expanded to form one or more lesions, and that subsequent driver gene mutations eventually led to invasive pancreatic cancer. We estimate that this multi-step progression generally spans many years. These new data reframe the step-wise progression model of pancreatic cancer by illustrating that independent, high-grade pancreatic precursor lesions observed in a single pancreas often represent a single neoplasm that has colonized the ductal system, accumulating spatial and genetic divergence over time.

The transformation of a normal cell to invasive cancer occurs through the accumulation of genetic and epigenetic changes[1]. Many invasive carcinomas in adults develop from morphologically recognizable noninvasive precursor lesions[2]. The most common precursor lesion associated with pancreatic ductal adenocarcinoma (PDAC) is pancreatic intraepithelial neoplasia (PanIN)[3]. At the morphological level, low-grade PanINs (LG-PanIN, PanIN-1 and PanIN-2) have minimal to moderate cytologic atypia and higher-grade PanINs (HG-PanIN, PanIN-3) have severe cytologic atypia. HG-PanINs exhibit morphological features that are thought to facilitate progression to an infiltrating carcinoma[4].

Aspects of this progression are supported by genetic studies[4–6], but fundamental questions about the development of PDAC remain[7]. The majority of PanINs (regardless of grade) harbour *KRAS* mutations; HG-PanINs and invasive carcinomas are more likely to contain additional driver gene alterations such as those in *TP53*, *CDKN2A*, and *SMAD4*. Moreover, PanINs and neighbouring PDACs often share many genetic alterations in both passenger and driver genes[8,9]. Collectively, these observations suggest that a subset of PDACs arises from adjacent PanINs, just as a colorectal carcinoma can arise from an underlying adenoma[10]. However, in individuals with multiple anatomically distinct PanINs[11], the biological and genetic relationships among these lesions, and their clinical significance, are not fully understood[12]. For instance, cancerization of the pancreatic ducts (an invasive cancer growing back into the duct system and simulating PanINs) by an established PDAC recapitulates lesions with histopathological features that are difficult to distinguish from those of bona fide PanIN precursor lesions[13]. Furthermore, the importance of non-invasive precursor lesions was recently challenged by a whole-genomic-sequence analysis of pancreatic cancers that proposed that tumorigenesis of pancreatic cancer is neither gradual nor slow[14]. We posited that a genomic evaluation of PDAC and matched co-evolving PanINs would provide additional insights into the biology of pancreatic cancer precursors and the dynamics of step-wise progression.

Evolutionary scenarios

Fig. 1a presents the conceptual framework underlying the interpretation of sequencing data generated from one PanIN and PDAC in the same patient, outlining three possible scenarios that in theory might be found. In the first scenario, the PanIN and the PDAC do not share any somatic mutations and arose independently. In the second scenario, the PanIN shares a subset of the somatic passenger and driver gene mutations with the PDAC, but the PDAC contains additional driver or passenger gene alterations not present in the PanIN. Scenario 2 presumes that a common ancestral cell underwent initiation and clonal expansion before seeding the PanIN and PDAC, but neither the common ancestral cell nor the founding PanIN cell had yet acquired all of the genetic events required to generate an invasive neoplasm. In the third scenario, the PanIN and the PDAC share some passenger mutations and all driver gene alterations, and the ancestral cell that seeded both the PDAC and PanIN had already acquired all of the alterations required to form a malignant cancer.

Whole-exome sequencing and phylogenetic analysis

To investigate the progression patterns of pancreatic carcinogenesis, we prospectively screened more than 100 resected pancreases from over a three-year interval to identify those samples in which at least one LG-PanIN (PanIN-2) or HG-PanIN (PanIN-3) was present in a region that was anatomically distinct and far removed from that of the PDAC (see Methods). We excluded any patient with a personal or family history

---

[1] The David M. Rubenstein Center for Pancreatic Cancer Research, Human Oncology and Pathogenesis Program, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [2] Department of Pathology, University of California, Davis, Sacramento, CA, USA. [3] The Sol Goldman Pancreatic Cancer Research Center, The Johns Hopkins University School of Medicine, Baltimore, MD, USA. [4] Canary Center for Cancer Early Detection, Department of Radiology, Stanford University School of Medicine, Palo Alto, CA, USA. [5] Program for Evolutionary Dynamics, Harvard University, Cambridge, MA, USA. [6] The Ludwig Center, The Johns Hopkins University School of Medicine, Baltimore, MD, USA. [7] VU University Amsterdam, Master's Oncology Program, VU University Medical Center, Amsterdam, The Netherlands. [8] Department of Cancer Genome Informatics, Graduate School of Medicine, Osaka University, Osaka, Japan. [9] Department of Art as Applied to Medicine, The Johns Hopkins University School of Medicine, Baltimore, MD, USA. [10] Department of Pathology, The Johns Hopkins University School of Medicine, Baltimore, MD, USA. [11] Department of Oncology, The Johns Hopkins University School of Medicine, Baltimore, MD, USA. [12] Department of Pathology, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [13] Department of Mathematics, Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA, USA. [14] Howard Hughes Medical Institute, The Johns Hopkins Kimmel Cancer Center, Baltimore, MD, USA. [15] These authors contributed equally: Alvin P. Makohon-Moore, Karen Matsukuma, Ming Zhang, Johannes G. Reiter, Jeffrey M. Gerold. *e-mail: iacobuzc@mskcc.org

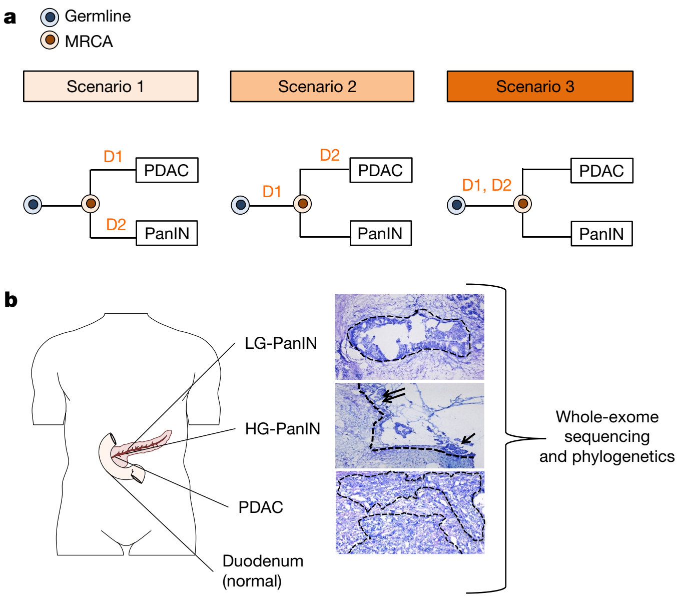


Fig. 1 | Evolutionary scenarios and study strategy of coexistent PanIN(s) and PDAC. a, Evolutionary scenarios of coexistent PanIN(s) and PDAC. For each of the three evolutionary scenarios, D1 and D2 indicate two hypothetical driver gene alterations; blue cells represent the germline (matched normal sample) and orange cells represent the most recent common ancestor (MRCA) for each PanIN–PDAC pair. In scenario 1, none of the somatic gene alterations are shared by the PanIN and PDAC. Mutation D1 is restricted to PDAC and mutation D2 is restricted to the PanIN. In scenario 2, only D1 is shared by the PanIN and PDAC. The mutation in D2 is restricted to the PDAC. In scenario 3, both D1 and D2 are shared by the PanIN and PDAC. **b**, Tissue collection, histological review and microdissection, WES, and phylogenetic analysis of patients. Body diagram adapted from the Motifolio toolkit. Micrographs show examples of PanINs and matched PDAC. The dashed outlines indicate regions that underwent laser capture microdissection for DNA extraction followed by whole-exome sequencing (WES). The LG-PanIN shows well formed papillary structures with nuclear crowding and cytological atypia. The HG-PanIN has regions of pseudopapillary formation (arrows) with a high nucleus–cytoplasm ratio. The matched PDAC shows features of poorly differentiated carcinoma with desmoplasia.

of PDAC from our study, as the dynamics of initiation in patients with germline alterations may be different from that in patients with sporadic pancreatic carcinogenesis[15]. Eight patients were identified, from which twelve PanINs and eight PDACs were sampled for the current study (Supplementary Table 1). All 20 tissue samples were laser-capture microdissected to ensure that a high fraction of the cells within each lesion were neoplastic (Fig. 1b). Despite the microscopic size of the PanINs, we were able to obtain sufficient amounts of DNA to generate high-quality libraries for whole-exome sequencing (WES). Notably, the generation of these libraries did not require whole-genome amplification before WES, thus reducing potential errors in downstream analyses.

We prepared sequencing libraries from each of the lesions as well as from normal tissues from each patient, and used them for massively parallel sequencing on an Illumina HiSeq instrument. We obtained a median canonical exon coverage of 253× across all samples. By comparing each lesion with its matched normal DNA, we identified 2,886 somatic single base substitutions (single nucleotide variations, SNVs) and small insertions or deletions (indels) (Extended Data Fig. 1, Supplementary Table 2). As a group, the PanINs harboured as many SNVs and indels as the PDACs (average of 75 and 80, respectively; Extended Data Fig. 1b). We also analysed somatic copy number alterations (CNAs) and structural variants from the exomic sequencing data (Supplementary Tables 3, 4, Extended Data Figs. 1c, 2). The number of CNAs, unlike the number of SNVs and indels, was higher in PDACs than in PanINs (average of 90 and 68, respectively).

Computational analysis (see Methods) revealed somatic mutations in many well-known driver genes, such as *KRAS*, *CDKN2A*, *TP53*, *SMAD4*, *U2AF1*, and *KMT2D* (Supplementary Table 5). Collectively, the genetic features of this set of PanINs and PDACs were consistent with the results of previous sequencing studies of these tumours[16–20]. To infer evolutionary relationships among the PanINs and PDACs for each patient based on the SNVs and indels, we used Treeomics[21], a phylogenetic method designed specifically for analysing sequencing data from spatially distinct tumours in the same individual[22] (see Methods). Treeomics identified high-confidence phylogenies for the matched samples from each of the eight patients (Fig. 2a–c, Extended Data Figs. 3–5). These analyses allowed us to derive the evolutionary relationships between the coexisting PanINs and the PDAC in each patient.

Evolutionary patterns in PDACs and PanIns

In our cohort, we found two patients (PIN102 and PIN105) in which no passenger gene mutations were shared by the PDAC and PanIN (Fig. 2a, Extended Data Fig. 3). For example, in patient PIN105 both the PanIN and PDAC had a KRAS(G12D) missense mutation. The PDAC exhibited 80 additional point mutations, including a one base-pair frameshift deletion in *TP53*, a missense mutation in *ACVR1B* (ACVR1B(C34Y)), and a 15-base-pair in-frame deletion in *SMAD4*. In addition, the PDAC had acquired CNA losses affecting *CDKN2A*, *MAP2K4*, *TP53*, and *SMAD4* (Extended Data Fig. 3b). The PDAC and PanIN may have arisen independently and by chance accumulated the same *KRAS* mutation (scenario 1), or they may have been initiated by a single KRAS(G12D) mutant clone and subsequently diverged (that is, scenario 2). Scenario 1 may be more likely, given the high frequency of *KRAS* variants in PDAC (more than 90%)[13] and the absence of any other shared somatic variants among the matched PanIN and PDAC samples in both of these patients. Moreover, the PanINs in both of these cases exhibited PanIN-2 histology, and a previous study indicated that LG-PanINs often harbour genetic features that support independent evolution. We note that the previous observation included distinct *KRAS* variants in matched PanINs, in contrast to the two cases presented here[9].

Four of the eight patients showed unequivocal evidence for scenario 2: a common ancestral cell underwent initiation and clonal expansion to form one or more PanINs. Further clonal expansions driven by additional driver gene mutations in a PanIN cell eventually led to a PDAC (Fig. 2b, Extended Data Fig. 4). For example, in patient PIN101, the common ancestor of PanIN lesion A and the PDAC acquired 14 somatic passenger mutations, including KRAS(G12D), as well as losses affecting *ACVR1B*, *MAP2K4*, *TP53*, and *SMAD4* (Fig. 2b, Extended Data Fig. 4a). The PDAC accumulated 28 point mutations, including CDKN2A(A21D) and TP53(R273H) missense mutations, a loss affecting *CDKN2A* and a gain affecting *MYC*. PanIN lesion A accumulated 111 point mutations, including a nonsense mutation in *SDK2*. Similar patterns were found in PIN103, PIN104 and PIN108—driver gene mutations common to all lesions as well as additional driver gene mutations specific to the PDAC (that is, scenario 2; Fig. 2b).

Finally, we observed two patients (PIN106 and PIN107) with phylogenetic patterns consistent with scenario 3, in which all lesions in a single pancreas shared all of the driver gene mutations identified (Fig. 2c, Extended Data Fig. 5). In patient PIN106, the common ancestor of all four samples harboured 47 somatic point mutations, including KRAS(G12D) and TP53(G266E) missense mutations, a mutation affecting the splice region in *ATM*, and GLI3(Q597*) (Fig. 2c). The PDAC subsequently acquired 39 passenger mutations and losses affecting *CDKN2A* and *SMAD4*.

In summary, the lesions in four of these eight patients were unequivocally derived from the same precursor clones, as they shared multiple passenger genes and a subset of driver genes (scenario 2). The presence of these additional driver gene alterations, coupled with phylogenetic analysis, provides persuasive evidence that the PDAC was derived from a PanIN in each case. These results highlight the value

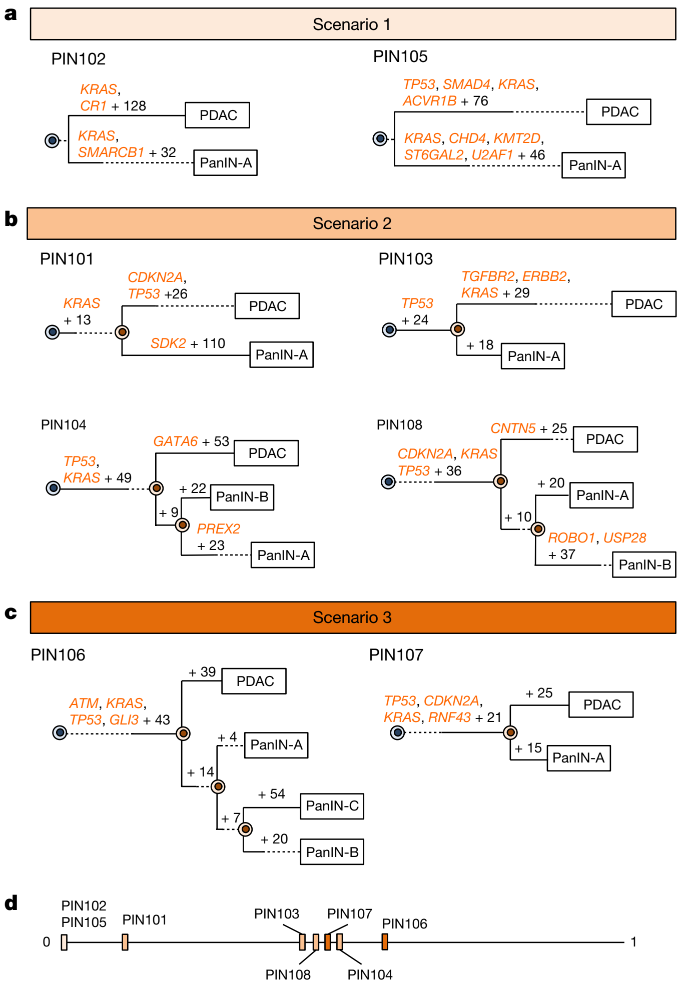


Fig. 2 | Phylogenetics of eight patients. a–c, Phylogenetic trees from SNVs and indels. The sample size is $n = 20$ lesions from eight patients. See Supplementary Table 1 for sample identities. For each phylogeny, gene names are SNVs or indels and the number of additionally acquired mutations is shown. The branch lengths approximate the number of SNVs or indels. The dashed lines indicate branches that have been extended to accommodate gene annotation and variant numbers. The sequencing data for each driver gene variant was manually reviewed to verify phylogenetic position. **a**, PIN102 and PIN105 are both scenario 1. **b**, PIN101, PIN103, PIN104, and PIN108 are scenario 2. In manual review of the PIN101 sequencing data, a read supporting the presence of the KRAS(G12D) variant was detected in both the PDAC and PanIN-A samples and was thus moved to the trunk of the phylogeny despite the overall low coverage of *KRAS* in PanIN-A. **c**, PIN106 and PIN107 are scenario 3. **d**, Jaccard indices from SNVs and indels. For each evolutionary scenario, the average Jaccard index for each patient was calculated from all driver and passenger variants (see Supplementary Table 6 for values) and plotted on a range from 0 to 1, with values closer to 1 denoting higher genetic similarity.

of genetic evaluation of morphologically distinct legions in revealing the evolutionary dynamics of pancreatic carcinogenesis. Because the PanINs were all anatomically distinct and far removed from the PDAC (see Methods), the data indicate that a single mutant clone had spread through the pancreatic ductal system to generate coexisting neoplastic lesions (Fig. 3a). This situation is similar to what occurs in the bladder, where a single clone can form multiple anatomically distinct neoplasms[23]. Though it would seem much more challenging for a neoplastic cell to travel through the fluid in the pancreatic ductal system than to travel through the urine, this journey has been described in intraductal papillary mucinous neoplasms of the pancreas, and clearly occurred in these four patients as well[24].

To assess genetic relatedness using all somatic variants, we quantified Jaccard similarity coefficients between pairs of lesions within each scenario (Fig. 2d, Supplementary Table 6). Notably, scenario 2 PanIN lesions tended to share fewer somatic variants with the matched PDAC than PanIN lesions in scenario 3 (average Jaccard similarity coefficient of 0.39 and 0.50, respectively), although the range of Jaccard similarity coefficients overlapped between the two scenarios (scenario 2 range, 0.10–0.57; scenario 3 range, 0.44–0.70).

Our phylogenetic analysis also enabled us to estimate the mutational signatures operating in different tumour lineages that led to the PDAC or a coexisting PanIN (Extended Data Figs. 6–8). Some signatures were shared between a PDAC and PanIN, whereas others operated only on a subset of different branches[25].

In PIN106 and PIN107, the PDACs and corresponding PanINs contained the same driver gene SNVs or indels (scenario 3; Fig. 2, Extended Data Fig. 5). In addition to the lost copies of *CDKN2A* and *SMAD4*, several unobserved factors might contribute to the morphological differences between lesions. First, the PDAC may have accumulated additional important genetic events in regions of the genome not assessed by whole-exome sequencing. Second, the PDACs may have acquired epigenetic alterations that were not detectable by the approach we used. Third, the microenvironment may have influenced the progression from a PanIN to a PDAC[13]. Finally, the PanIN lesions in PIN106 and PIN107 may represent cancerization of the ducts as described above. We note that the PDACs in these two patients showed moderate to poorly differentiated histology, thereby decreasing but not fully eliminating this possibility[12].

Modelling time of pancreatic cancer evolution

The WES data allow us to estimate the time required for a cell to progress from a non-invasive, neoplastic clone to an invasive pancreatic cancer[26] (see Methods). We used the number of acquired genetic passenger alterations from a common ancestor to the PanINs and the PDACs, after removing mutations suspected to be drivers or subclonal, to infer the amount of passed time. Because the great majority of the mutations present in any of these lesions are passengers and are not associated with positive or negative growth advantage, these mutations can serve as a molecular clock. On the basis of previously estimated mutation rates[27] and cell division times[28] measured in PanINs, we found that the median time elapsed between the common ancestral cell (Fig. 3b) and the birth of the founder clone of a PanIN was 7.1 years (90% confidence interval (CI) of the median: 3.3–12.2 years). Similarly, the median time elapsed between the common ancestral cell and the founder cell of the PDAC itself was 4.3 years (90% CI 2.3–7.2 years). Because the PanIN samples are monophyletic in all patients, we cannot estimate how long the primary tumour lineage might have existed as a PanIN. Nonetheless, these intervals are conservative underestimates of the times required to develop neoplasia and radiographically detectable cancer, because they do not include any clonal steps before the birth of the common ancestral cell nor the time between the birth of the PDAC founder cell and the multiplication of this cell to form a clinically evident mass. A larger patient cohort is required to assess whether this length of time is characteristic of the population of individuals with PDAC. When the time required for mass development is taken into account, the data suggest that an average of at least 8.1 years elapses between the birth of the common ancestral cell and the presence of a clinically evident mass (see Methods).

Discussion

Comparison of our results with three recent studies is informative. First, 77% of patients without clinically evident pancreatic neoplasia were found to harbour PanIN-1 lesions at autopsy[11]. In another study, LG-PanINs (PanIN-1 and PanIN-2) from the same patient generally did not share the same genetic alterations, in contrast to our data which show genetic relationships among HG-PanINs (PanIN-3)[9]. When taken together with our results, the data suggest that early neoplastic lesions in the pancreas may represent independent events, and that the success of the neoplastic cells in colonizing the ductal system is achieved only with histological progression and the accrual of additional genetic

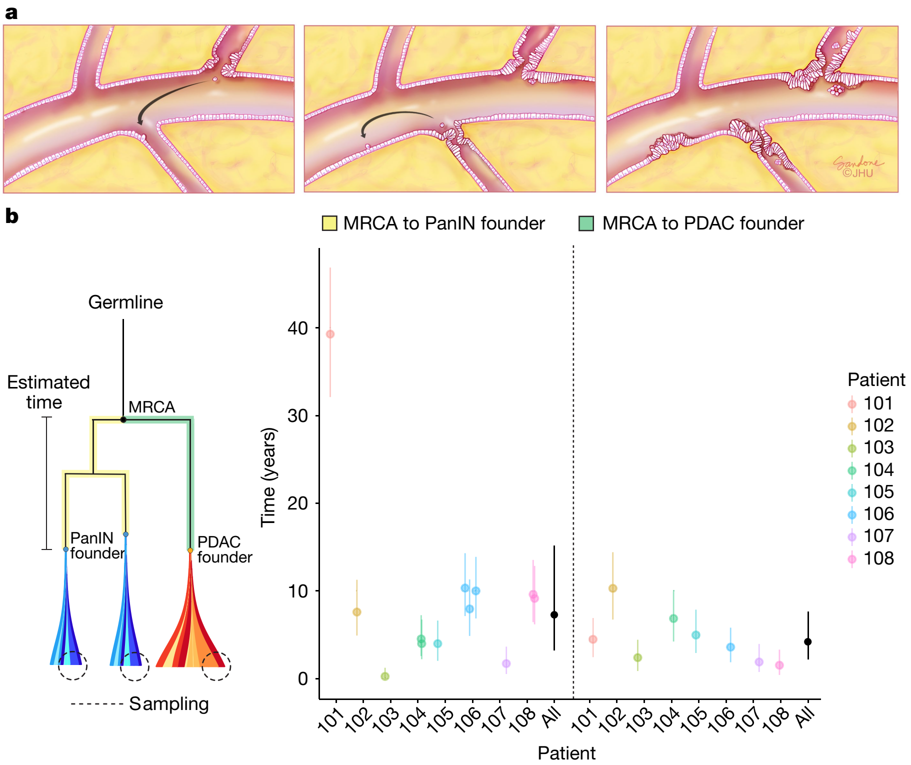


Fig. 3 | Putative growth pattern of coexisting PanIN(s) and PDAC and mathematical model. a, Spatial evolution and PanIN progression in intralobular ducts. LG-PanIN and HG-PanIN lesions represent precursors with differing degrees of nuclear and cytological atypia. An LG-PanIN develops and seeds a cell that travels to a second duct (arrow, left). The first LG-PanIN matures into an HG-PanIN, while an LG-PanIN develops at the second site and a cell subsequently travels to a third duct (arrow, middle). The second site LG-PanIN matures into an HG-PanIN while an LG-PanIN develops at the third site (right). **b**, Estimated progression times. The lineage leading from the MRCA to the PanINs is illustrated in yellow, while the lineage leading from the MRCA to the PDAC is in green. Clonal passenger mutations were used to estimate progression times, shown for each patient with 90% CIs. Overall (black), the inferred median time elapsed between the common ancestral cell and the birth of the founder clone of a PanIN was 7.1 years (90% CI, 3.3–12.2; MRCA to PanIN, $n = 12$). The inferred median time elapsed between the common ancestral cell and the PDAC was 4.3 years (90% CI, 2.3–7.2; MRCA to PDAC, $n = 8$). These estimates assume a mutation rate of 0.0224 per generation and a time per generation of four days (see Methods).

alterations. Notably, the budding off of small clusters of neoplastic epithelial cells into the lumen is one of the pathognomonic morphological features of an HG-PanIN (PanIN-3)[29].

Our data are apparently at odds with the interpretation of a recent study that concluded that PDACs do not arise gradually[14]. This conclusion was based on genetic analyses of microdissected PDACs and did not include an analysis of PanINs, and models were not applied to the data to support such a conclusion. As such, it relied on assumptions about the timing of transition from precursor lesion to invasive carcinoma. By contrast, our data are directly based on genomic analyses of the precursor lesions and their corresponding PDACs. Our step-wise progression model is supported not only by the current data but also by a body of scientific literature[17–20,22,26,30,31] that suggests that single or short base substitutions that gradually accumulate over many years form the great majority of the genetic alterations responsible for this type of tumour. Our findings in no way contradict the observation that multiple chromosome translocations can occur simultaneously (chromothripsis) in a small subset of pancreatic tumours[14,31]. However, they do buttress the idea that PDAC development is a multi-step progression caused by the accumulation of somatic alterations in driver genes, in a process that generally spans many years.

It could be argued that the cases we analysed were unusual in that more than one advanced PanIN was found in each pancreas, and our selection of eight out of approximately one-hundred patients could have introduced an unintended bias in our cohort. However, multiple advanced PanIN lesions are the norm rather than the exception when the entire pancreas is methodically dissected[11]. Furthermore, the mutations in driver genes and distribution of mutational signatures in this cohort are similar to those previously observed in pancreatic cancers. Finally, genomic analysis of a PDAC arising directly from an adjacent HG-PanIN lesion revealed a gradual genetic progression from PanIN to PDAC[8]—similar to our findings for anatomically separate HG-PanIN lesions and their corresponding PDACs.

In summary, we have shown that pancreatic intraepithelial neoplasia (that is, PanIN-2 and PanIN-3) need not be spatially localized lesions; rather, they are a disease that can spread through the entire ductal system. Additional studies—with more patients and a higher density of samples—will be required to determine the frequencies of the evolutionary scenarios we identified and to clarify which features of precursor lesions put them at substantial risk of transformation. Nonetheless, our data suggest that the multiple, apparently discrete PanIN lesions observed in an individual patient often represent a single neoplasm that can spread (contiguously or discontiguously) along the ductal system. This finding provides an explanation for the observation that patients who have had an HG-PanIN or PDAC removed by subtotal pancreatectomy are at high risk for the development of recurrent disease.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0481-8

Received: 1 February 2018; Accepted: 13 July 2018;
Published online 3 September 2018.

1. Vogelstein, B. et al. Cancer genome landscapes. *Science* **339**, 1546–1558 (2013).

2. Vogelstein, B. & Kinzler, K. W. The path to cancer —three strikes and you're out. *N. Engl. J. Med.* **373**, 1895–1898 (2015).
3. Basturk, O. et al. A revised classification system and recommendations from the Baltimore consensus meeting for neoplastic precursor lesions in the pancreas. *Am. J. Surg. Pathol.* **39**, 1730–1741 (2015).
4. Hruban, R. H., Goggins, M., Parsons, J. & Kern, S. E. Progression model for pancreatic cancer. *Clin. Cancer Res.* **6**, 2969–2972 (2000).
5. van Heek, N. T. et al. Telomere shortening is nearly universal in pancreatic intraepithelial neoplasia. *Am. J. Pathol.* **161**, 1541–1547 (2002).
6. Kanda, M. et al. Presence of somatic mutations in most early-stage pancreatic intraepithelial neoplasia. *Gastroenterology* **142**, 730–733.e9 (2012).
7. Makohon-Moore, A. & Iacobuzio-Donahue, C. A. Pancreatic cancer biology and genetics from an evolutionary perspective. *Nat. Rev. Cancer* **16**, 553–565 (2016).
8. Murphy, S. J. et al. Genetic alterations associated with progression from pancreatic intraepithelial neoplasia to invasive pancreatic tumor. *Gastroenterology* **145**, 1098–1109.e1 (2013).
9. Hosoda, W. et al. Genetic analyses of isolated high-grade pancreatic intraepithelial neoplasia (HG-PanIN) reveal paucity of alterations in *TP53* and *SMAD4*. *J. Pathol.* **242**, 16–23 (2017).
10. Fearon, E. R. & Vogelstein, B. A genetic model for colorectal tumorigenesis. *Cell* **61**, 759–767 (1990).
11. Matsuda, Y. et al. The prevalence and clinicopathological characteristics of high-grade pancreatic intraepithelial neoplasia: autopsy study evaluating the entire pancreatic parenchyma. *Pancreas* **46**, 658–664 (2017).
12. Yamasaki, S., Suda, K., Nobukawa, B. & Sonoue, H. Intraductal spread of pancreatic cancer. Clinicopathologic study of 54 pancreatectomized patients. *Pancreatology* **2**, 407–412 (2002).
13. Kleeff, J. et al. Pancreatic cancer. *Nat. Rev. Dis. Primers* **2**, 16022 (2016).
14. Notta, F. et al. A renewed model of pancreatic cancer evolution based on genomic rearrangement patterns. *Nature* **538**, 378–382 (2016).
15. Roberts, N. J. et al. Whole genome sequencing defines the genetic heterogeneity of familial pancreatic cancer. *Cancer Discov.* **6**, 166–175 (2016).
16. Jones, S. et al. Comparative lesion sequencing provides insights into tumor evolution. *Proc. Natl Acad. Sci. USA* **105**, 4283–4288 (2008).
17. Biankin, A. V. et al. Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes. *Nature* **491**, 399–405 (2012).
18. Waddell, N. et al. Whole genomes redefine the mutational landscape of pancreatic cancer. *Nature* **518**, 495–501 (2015).
19. Witkiewicz, A. K. et al. Whole-exome sequencing of pancreatic cancer defines genetic diversity and therapeutic targets. *Nat. Commun.* **6**, 6744 (2015).
20. Bailey, P. et al. Genomic analyses identify molecular subtypes of pancreatic cancer. *Nature* **531**, 47–52 (2016).
21. Reiter, J. G. et al. Reconstructing metastatic seeding patterns of human cancers. *Nat. Commun.* **8**, 14114 (2017).
22. Makohon-Moore, A. P. et al. Limited heterogeneity of known driver gene mutations among the metastases of individual patients with pancreatic cancer. *Nat. Genet.* **49**, 358–366 (2017).
23. Sanli, O. et al. Bladder cancer. *Nat. Rev. Dis. Primers* **3**, 17022 (2017).
24. Pea, A. et al. Targeted DNA sequencing reveals patterns of local progression in the pancreatic remnant following resection of intraductal papillary mucinous neoplasm (IPMN) of the pancreas. *Ann. Surg.* **266**, 133–141 (2017).
25. Roerink, S. F. et al. Intra-tumour diversification in colorectal cancer at the single-cell level. *Nature* **556**, 457–462 (2018).
26. Yachida, S. et al. Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature* **467**, 1114–1117 (2010).
27. Tomasetti, C., Vogelstein, B. & Parmigiani, G. Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation. *Proc. Natl Acad. Sci. USA* **110**, 1999–2004 (2013).
28. Klein, W. M., Hruban, R. H., Klein-Szanto, A. J. P. & Wilentz, R. E. Direct correlation between proliferative activity and dysplasia in pancreatic intraepithelial neoplasia (PanIN): additional evidence for a recently proposed model of progression. *Mod. Pathol.* **15**, 441–447 (2002).
29. Hruban, R. H. et al. An illustrated consensus on the classification of pancreatic intraepithelial neoplasia and intraductal papillary mucinous neoplasms. *Am. J. Surg. Pathol.* **28**, 977–987 (2004).
30. Jones, S. et al. Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science* **321**, 1801–1806 (2008).
31. Reiter, J. G. & Iacobuzio-Donahue, C. A. Pancreatic cancer: Pancreatic carcinogenesis—several small steps or one giant leap? *Nat. Rev. Gastroenterol. Hepatol.* **14**, 7–8 (2016).

Acknowledgements Supported by the V Foundation for Cancer Research, NIH grants F31 CA180682, 2T32 CA160001-06 and 5T32 CA067751-13, an Erwin Schrödinger fellowship (Austrian Science Fund FWF J-3996), SPORE grant P50 CA062924, the Michael Rolfe Foundation, The Lustgarten Foundation for Cancer Research, the Sol Goldman Center for Pancreatic Cancer Research, The Virginia and D.K. Ludwig Fund for Cancer Research and D. Troper and S. Wojcicki.

Reviewer information *Nature* thanks A. V. Biankin, S. J. Chanock and F. Markowetz for their contribution to the peer review of this work.

Author contributions A.P.M.-M., K.M., Y.J., N.P., K.K., B.V., and C.I.-D. designed the study; K.M., S.Y., R.H.H., and C.I.-D. selected the samples; S.Y., K.M., R.H.H., D.S.K., and C.I.-D. reviewed pathology; S.Y., K.M. and M.Z. prepared the DNA samples; A.P.M.-M., K.M., M.Z., Y.J., N.P., K.K., B.V., and C.I.-D. performed sequencing, alignment and mutation calling; A.P.M.-M., J.G.R., J.M.G., and M.A. derived the phylogenies; A.P.M.-M., J.G.R., J.M.G., M.A., and L.S. analysed the structural variants; J.G.R., J.M.G., and M.A.N. performed mathematical modelling; C.S. illustrated the spatial evolution of the lesions; and all authors wrote the manuscript.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0481-8.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0481-8.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to C.I.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

Patient selection. Human tissues were collected with the approval of the Johns Hopkins Hospital Institutional Review Board (protocols NA_00001584 and NA_00017879) after informed and written consent was obtained, following all relevant ethical regulations. Fresh-frozen samples from eight patients who underwent surgical resection of pancreatic cancer at Johns Hopkins Hospital (January 2009–December 2011) with pathological confirmation of PDAC and geographically distinct PanIN-2 or PanIN-3 lesions were selected for study. For inclusion in the study, PDAC, PanIN lesion(s), and normal duodenum tissue were required for each patient. To minimize the possibility of studying cancerization of normal ducts, we included only PanINs in which at least 1.0 cm of uninvolved lobular parenchyma was present between the PanIN and the cancer, or in which the PanINs were present in a block that contained no cancer.

Processing of tissue samples. For each tissue sample, multiple sequential 5-μm-thick cryosections were mounted on polyethylene naphthalate (PEN) membrane slides and stained with cresyl violet for visualization of histological features and confirmation of adequate cellularity. Neoplastic epithelium was laser-microdissected using the Leica LMD7 laser microdissection system.

DNA extraction and quantification. Genomic DNA (gDNA) was extracted from each normal, PanIN, or tumour piece using a standard phenol and chloroform extraction followed by precipitation in ethanol. The gDNA was quantified by LINE assay (counting long interspersed elements (LINEs) using real-time PCR). The LINE forward primer was 5′-AAAGCCGCTCAACTACATGG-3′ and the reverse primer was 5′-TGCTTTGAATGCGTCCCAGAG-3′. The real-time PCR protocol was 50 °C for 2 min, 95 °C for 2 min, 40 cycles of 94 °C for 10 s, 58 °C for 15 s, and 70 °C for 30 s, 95 °C for 15 s, and 60 °C for 30 s. The PCR reactions were carried out using Platinum SYBR Green qPCR mastermix (Invitrogen).

WES and alignment. WES was performed on an Illumina HiSeq 2000 platform for a target coverage of 150×. Upon the completion of WES, the data were analysed in silico to determine overall quality and coverage. Sequencing reads were aligned to the hg19 human reference genome using BWA[32]. Read de-duplication, base quality recalibration, and multiple sequence realignment were performed using the Picard Suite and GATK version 3.1[33,34]. SNVs were called using Mutect version 1.1.6 and indels were detected using HaplotypeCaller version 2.4[33,35].

Filtering of WES data. WES generated a large list of potential mutations, and we evaluated these data to identify high-quality mutations while removing sequence artefacts. Each mutant must have been observed with at least 5% variant allele frequency with 20× coverage in at least one neoplastic sample; each mutant must have been observed in less than 2% of the reads (or three reads total) of the matched normal sample with 10× coverage. This filtering yielded a total of 2,886 mutations for subsequent analysis (Supplementary Table 2).

Driver gene and mutation analysis. All somatic variants causing a frameshift deletion, frameshift insertion, in-frame deletion, in-frame insertion, missense, nonsense, nonstop or splice site/region mutation, or a translation start site were considered. If a variant was a missense or nonsense mutation, we required the variant to have a CHASM P value ≤ 0.05 and an FDR ≤ 0.25. In combination with manual review, driver gene mutations were identified if the gene was supported by at least three of the following four methods: 20/20+[36], TUSON[37], MutSigCV[38] (see Supplementary Table 1 in ref. 36 for gene list), and a hotspot analysis[39]. In addition, we also considered genes significantly mutated in large PDAC sequencing studies[17,18,20,40]. Further, we required that each somatic variant have a variant allele frequency of <2% in the patient-matched normal tissue as well as any normal tissue from another patient. If a deleterious variant was detected in a driver gene as described above, and was not detected abundantly in any normal tissue, it was considered a driver gene variant.

CNAs. Allele-specific copy number analysis was performed using FACETS[41]. In brief, FACETS performs a complete analysis that includes library size and GC-normalization, and segmentation of total and allele-specific signals, using coverage and genotypes of single nucleotide polymorphisms simultaneously across the exome[41]. The resulting segments accurately identify points of change in the exome, accounting for diploidy, purity, and average ploidy for each sample. A maximum likelihood approach then assigns each segment with a major and minor integer copy number.

Evolutionary analysis. We derived phylogenies for each set of samples using Treeomics 1.7.9[21]. Each phylogeny was rooted at the matched patient's normal sample and the leaves represented the PanIN or tumour samples. Treeomics employs a Bayesian inference model to account for error-prone sequencing and varying neoplastic cell content to calculate the probability that a specific variant is present or absent. Treeomics infers the global optimal tree based on Mixed Integer Linear programming. For Extended Data Figs. 3–5, the CNAs were not directly used to infer phylogenies in order to prevent bias from potential false-negatives or false-positives, given that CNA calls from multiple samples within a patient are particularly sensitive to varying neoplastic cell content and depth of sequencing. Moreover, WES data usually do not capture the exact breakpoints of CNAs, further complicating phylogenetic analysis. Nevertheless, the common PDAC driver genes *KRAS*, *MYC*, *GATA6*, and *CDK6* were manually reviewed in the CNA data for evidence of gains, while *CDKN2A*, *SMAD4*, *TP53*, *MAP2K4*, *TGFBR2*, and *ACVR1B* were queried for losses. Allelic losses were defined as total copy number (tcn) = 1 or 0, and gains were defined as tcn ≥ 4. Given the CNA status of a given driver gene in each sample, the driver gene with the CNA status was manually placed on the corresponding position edge in the phylogeny (previously derived using SNVs/indels). This approach was used with each PDAC driver gene affected by a CNA as defined above.

Our classification of each patient into one of three evolutionary scenarios was based on SNVs/indels that affect key driver genes in PDAC (for example, KRAS(G12D)). Such alterations represent driver gene variants that are readily interpretable with respect to function as well as position on the phylogenetic tree. Nonetheless, CNAs can also affect driver genes involved in pancreatic cancer (for example, *CDKN2A* deletion). If we reclassify the eight patients using both SNVs/indels and CNAs affecting driver genes (Extended Data Figs. 3–5), we find that the evolutionary scenario does not change for six patients. For two patients (PIN106 and PIN107), the scenario changes from scenario 3 to scenario 2, indicating a step-wise progression of PanINs and PDACs for all eight patients. As noted above, the identification and placement of CNAs on a phylogenetic tree remains challenging. Nonetheless, we note that the SNV/indel phylogenies represent a minimum number of evolutionary steps: including additional CNAs would either confirm or increase the total number of steps in the evolution of the PDAC.

Structural variant analysis. We inferred structural variants (SVs) using DELLY2 (v.0.7.5) to verify the reconstructed phylogenies[42]. As the SVs were called for each sample independently, we merged SVs for which DELLY determined breakpoints differing by at most 250 base pairs among the samples of each patient. In total, we found 154 distinct SVs in the eight subjects. After a comprehensive manual review of the called SVs, we developed additional criteria to minimize the number of false-positives. We required that each SV had to pass one of the following two filters in at least one sample: 1) (a) SV is supported by at least three distinct split reads, (b) the ratio of split reads that support the SV to the total number of split reads at the position of the SV is greater or equal to 0.75, and (c) the number of the SV supporting split reads is greater than the number of split reads in the normal sample; or 2) (d) the SV is supported by at least five discordantly paired (DP) reads, (e) the ratio of DP reads that support the SV to the total number of DP reads at the position of the SV is greater or equal to 0.25, and (f) the number of the SV-supporting DP reads is greater than the number of DP reads in the normal sample. After applying these filters, we obtained 40 SVs (Supplementary Table 4).

To create input files for Treeomics, we used the number of SV-supporting split and DP reads as the number of variant reads. We normalized the coverage of SVs such that on average it approximately matched the median coverage of SNVs. Generally, the inferred phylogenies based on the SVs agreed well with the ones based on SNVs. However, because of the lower number of SVs per subject (median 4; range 0–14; Supplementary Table 4), the confidence in the inferred branches was substantially lower than in the phylogenies based on SNVs. For PIN106 (coverage in sample of PanIN-A was extremely low), we inferred a slightly different phylogeny as PanIN-A diverged before the PDAC, probably owing to many false-negatives resulting from the extremely low coverage and the therefore difficult detection of SVs in this sample. For PIN108, no SVs were shared across multiple samples and hence there were no parsimony-informative SVs such that a phylogeny could be inferred.

Mutation signatures. We assessed the presence of previously identified mutational signatures[43] in each patient. Our phylogenetic analysis enabled us to estimate the signatures operating at different stages of cancer evolution[25]. For SNVs acquired along each phylogenetic branch, we estimated the maximum likelihood signature proportions among 30 previously identified trinucleotide signatures[44] (see https://github.com/mskcc/mutation-signatures). We quantified the uncertainty in these estimates by performing 100 iterations of bootstrap resampling within each branch followed by signature re-estimation. We ignored branches with five or fewer mutations and removed signature 24 because of its similarity to smoking. The maximum likelihood signature estimates and 90% bootstrap confidence intervals for each branch are shown in Extended Data Figs. 6–8. We detect signatures 1, 2, 3, and 6, consistent with previous studies[43]. Additionally, we find evidence for signatures 4 (associated with smoking) and 29 (associated with chewing tobacco). Signatures operating on different branches within a patient were not significantly more similar than those across patients (mean cosine distance similarity 0.62 versus 0.59, $P = 0.21$, one-sided permutation test). We note that signature estimates had large bootstrap uncertainty and the number of patients as well as the number of mutations is limited.

Progression time inference. We assume that the number of passenger mutations n acquired along a lineage during time T (in cell generations) is Poisson-distributed with rate equal to T times the mutation rate per cell division[16]: $n_\mu \mid T \sim$ Poisson (μT). We assume that a random sample from the population of PanINs or PDACs takes T generations to progress from a previous stage (either MRCA of all sampled PanINs and PDAC in a patient or the MRCA of the most closely related PanIN to the PDAC) to the founder of a particular PanIN or PDAC, and that the mutational clock time μT is gamma-distributed with hyperparameters shape k and scale θ ($k, \theta > 0$) uniform a priori: $\mu T \sim$ Gamma (k, θ).

In order to infer the joint distribution of (T, k, θ), we use the following sampling strategy. For each sample i, we update T by sampling directly from the gamma posterior:

$$T_i\mu|n, k, \theta \sim \text{Gamma}\left(k+n, \frac{\theta}{1+\theta}\right)$$

Using the updated values, we jointly update k, θ by Metropolis–Hastings sampling from the posterior:

$$L(k, \theta) \propto \pi(k, \theta) \prod_{T_i} d\text{Gamma}(k, \theta, \mu T_i)$$

where dGamma is the density function for the gamma distribution and $\pi(k,\theta)$ is the prior over the hyperparameters (uniform). This setup pools information about the time to progression for each sample towards the population of progression time estimates, with a flexible structure for the overall distribution of times provided by the gamma distribution.

In order to convert the inferred number of generations to absolute time, we follow a previous method[26] by multiplying by the average time for cell division. To estimate the division time, we again follow the previous method but instead note that 14% of stage II PanINs stain positively for Ki-67[28]. We therefore estimate the generation time of PanIN stage II cells to be 4 days. The mutation rate μ per generation is 0.0224, calculated for 35 Mb of exome sequencing multiplied by a point mutation rate of 6.4×10^{-10} per generation[27].

To calculate the expected time it takes for the PDAC founding cell to grow to a detectable lesion of 1 cm^3 ($\sim 10^9$ cells), we used previously measured PDAC metastasis doubling times of 56 days[45] leading to an exponential growth rate of $r = 0.012$ per day. The probability density function for the time an exponential branching process conditioned on survival takes to reach size $M = 10^9$ is approximately given by:

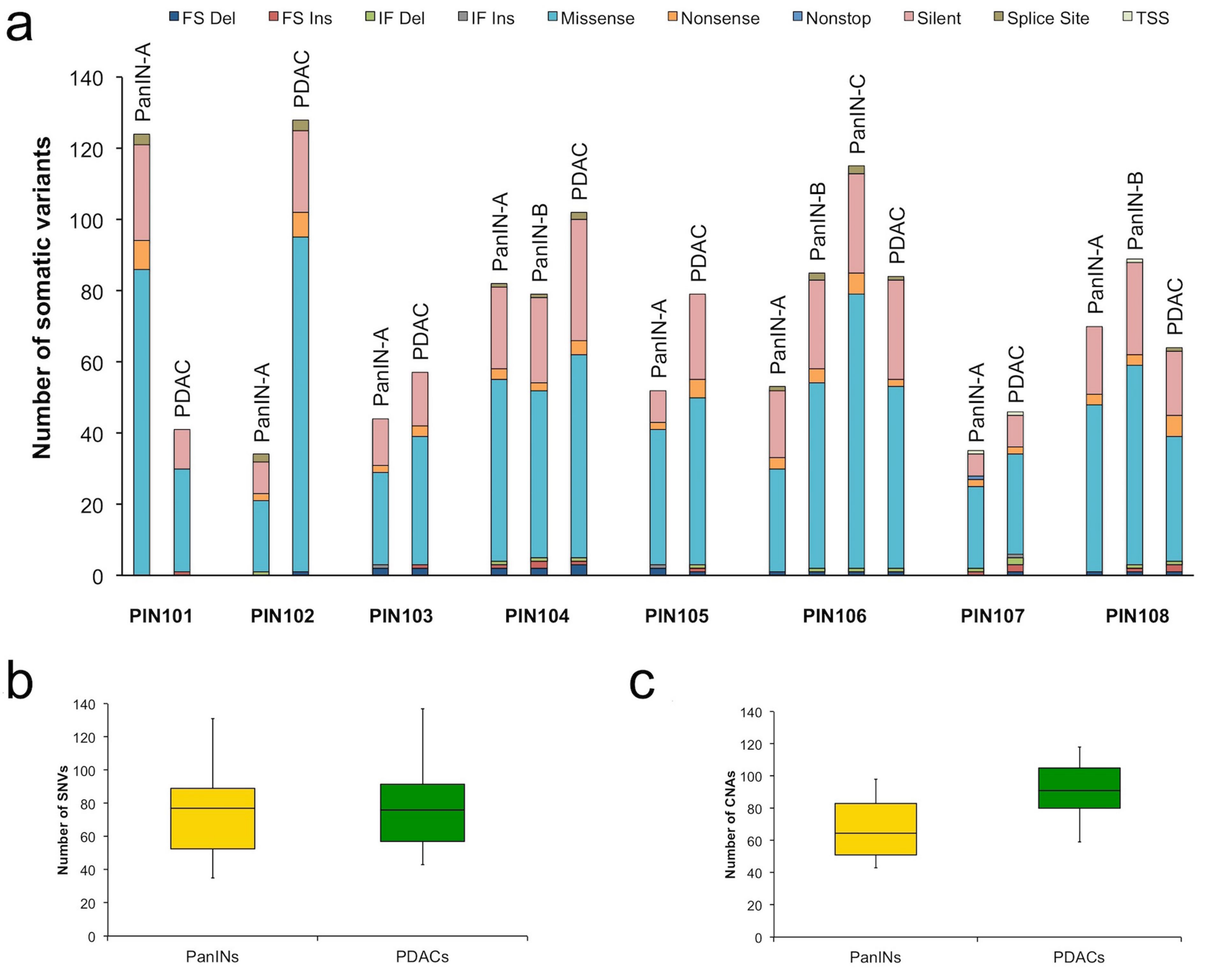
$$f_{t_M}(t) = \exp\left(-\frac{r}{b} M \exp(-rt)\right) \frac{r^2 M}{b} \exp(-rt)$$

Where $b = 1/2.3$ per day is the assumed PDAC cell division rate[26,46].

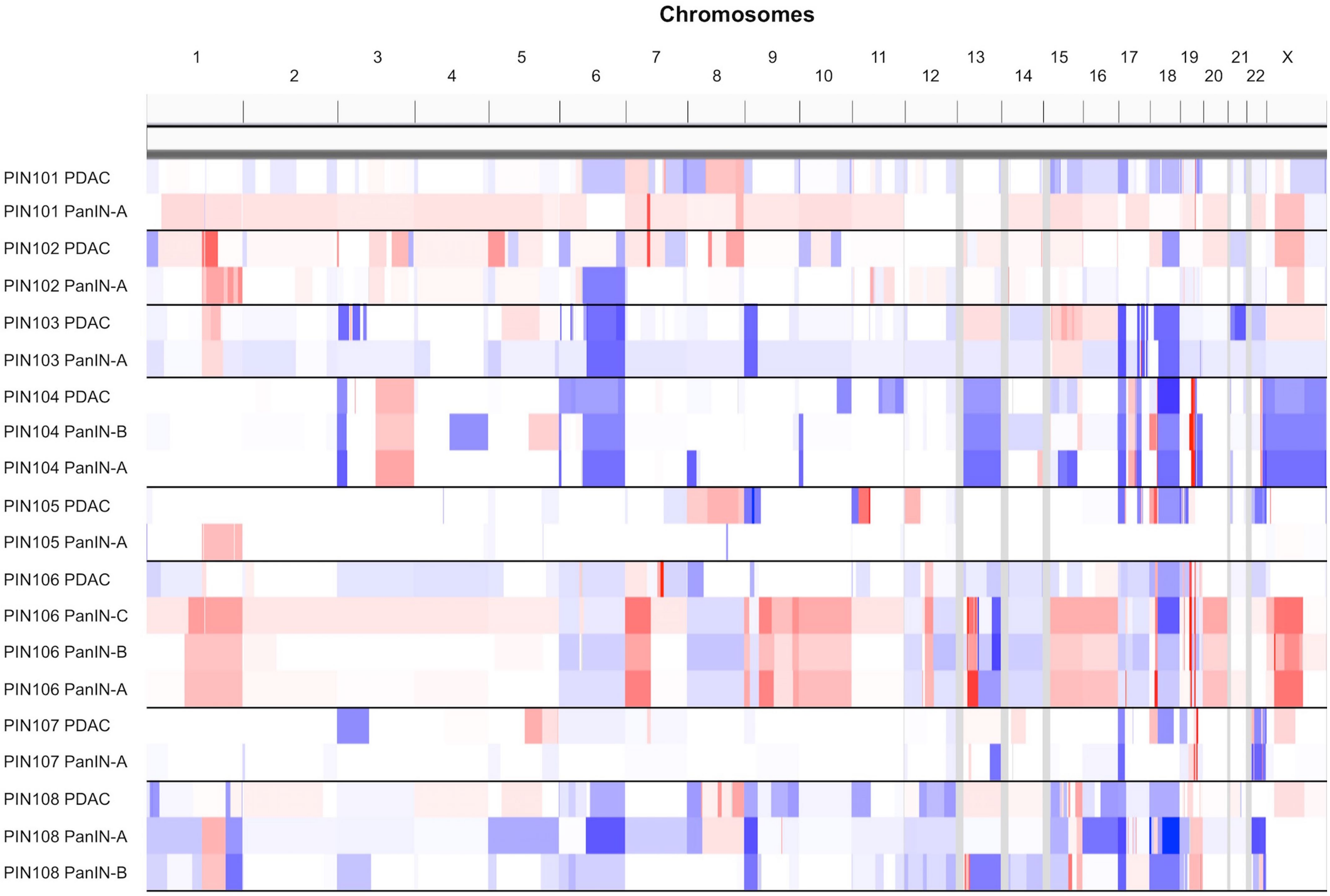
Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability. Sequence data have been deposited at the European Genome-phenome Archive (EGA; https://www.ebi.ac.uk/ega/), which is hosted by the European Bioinformatics Institute (EBI) and the Centre for Genomic Regulation (CRG), under accession number EGAS00001002778. Source data are provided for Fig. 3b and Extended Data Figs. 1, 7, 8. All other relevant data are included within the manuscript or are available upon request from the corresponding author.

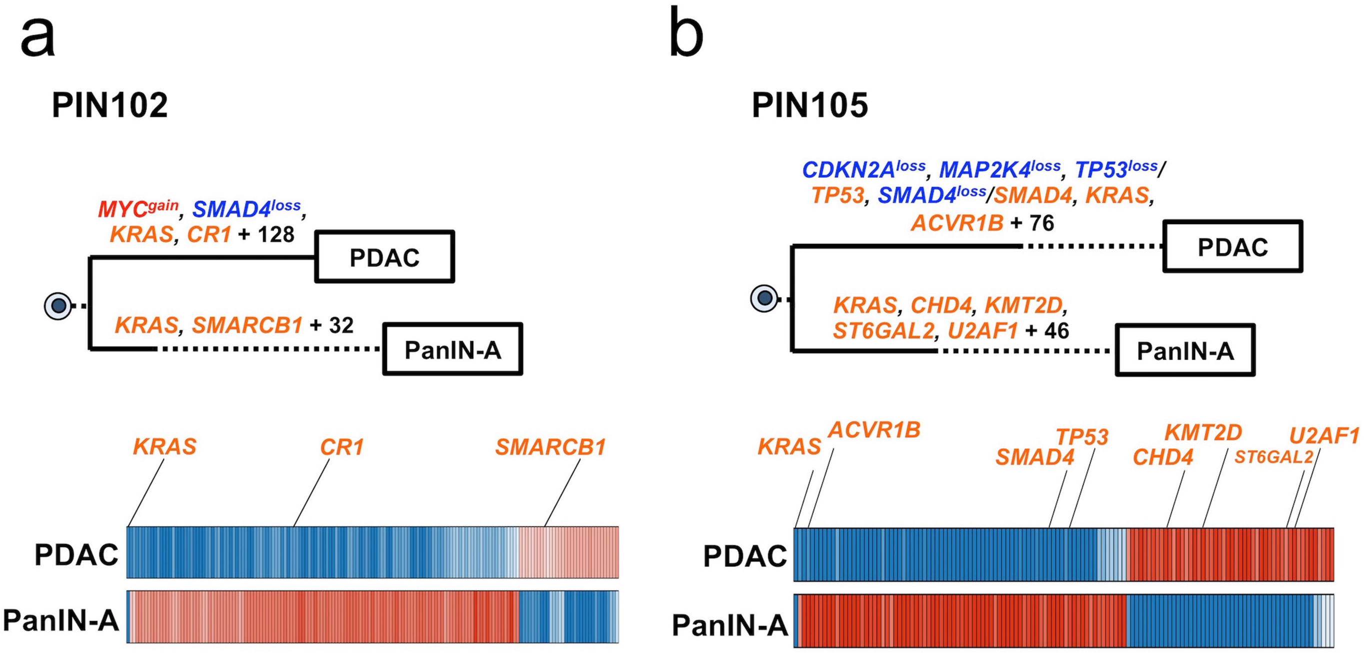
32. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
33. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
34. Mose, L. E., Wilkerson, M. D., Hayes, D. N., Perou, C. M. & Parker, J. S. ABRA: improved coding indel detection via assembly-based realignment. *Bioinformatics* **30**, 2813–2815 (2014).
35. Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).
36. Tokheim, C. J., Papadopoulos, N., Kinzler, K. W., Vogelstein, B. & Karchin, R. Evaluating the evaluation of cancer driver genes. *Proc. Natl Acad. Sci. USA* **113**, 14330–14335 (2016).
37. Davoli, T. et al. Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell* **155**, 948–962 (2013).
38. Lawrence, M. S. et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
39. Chang, M. T. et al. Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. *Nat. Biotechnol.* **34**, 155–163 (2016).
40. Cancer Genome Atlas Research Network Integrated genomic characterization of pancreatic ductal adenocarcinoma. *Cancer Cell* **32**, 185–203.e13 (2017).
41. Shen, R. & Seshan, V. E. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. *Nucleic Acids Res.* **44**, e131 (2016).
42. Rausch, T. et al. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333–i339 (2012).
43. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
44. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. Deciphering signatures of mutational processes operative in human cancer. *Cell Reports* **3**, 246–259 (2013).
45. Amikura, K., Kobari, M. & Matsuno, S. The time of occurrence of liver metastasis in carcinoma of the pancreas. *Int. J. Pancreatol.* **17**, 139–146 (1995).
46. Durrett, R. in *Branching Process Models of Cancer* 1–63 (Springer International Publishing, Switzerland, 2015).



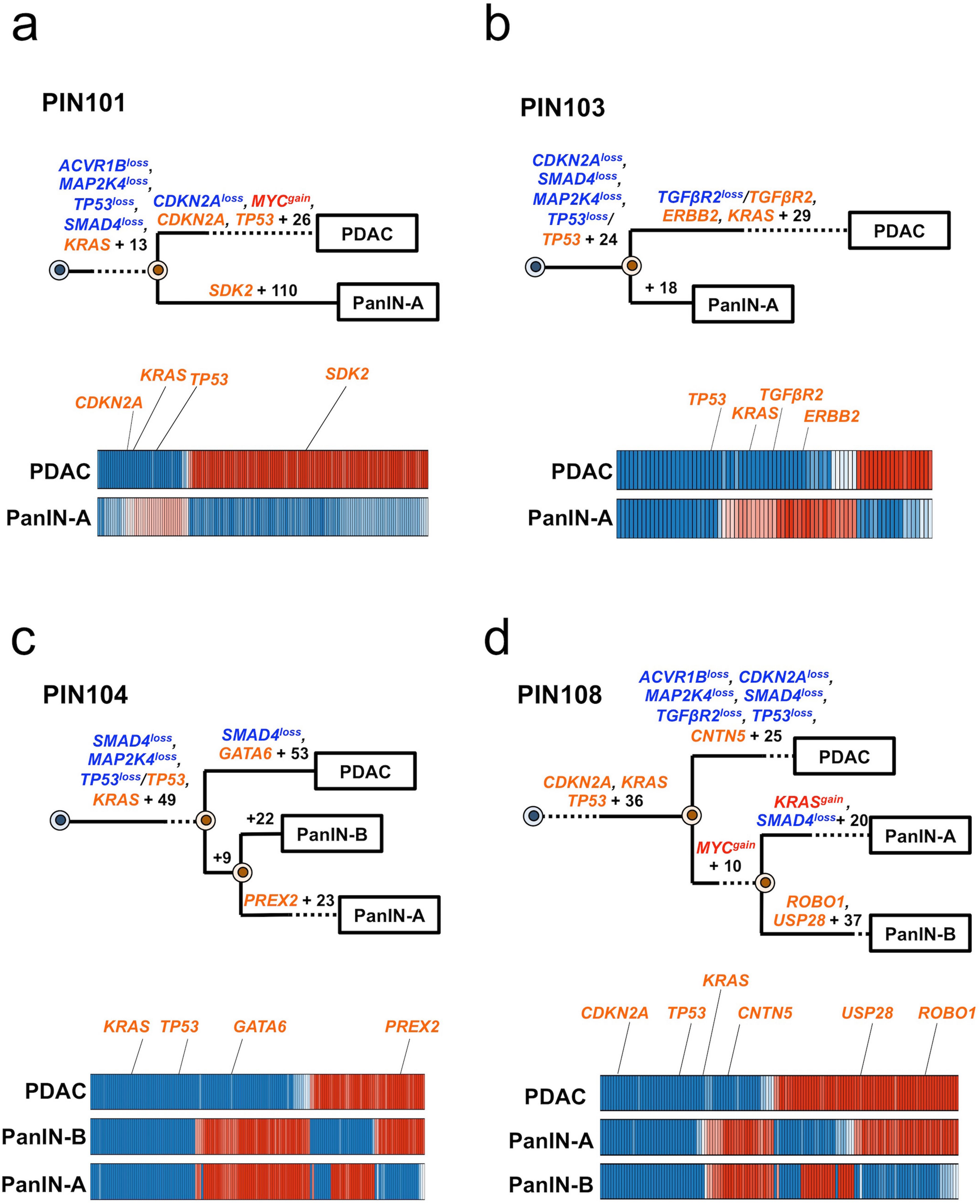
Extended Data Fig. 1 | Mutation counts and features of samples. **a**, Number of somatic mutations detected per sample with clinical features of each patient. The *y*-axis shows mutation counts and the *x*-axis shows patient ID. FS Del, frameshift deletion; FS Ins, frameshift insertion; IF Del, in-frame deletion; IF Ins, in-frame insertion; TSS, transcription start site. **b**, **c**, Box and whisker plots comparing number of somatic SNVs and CNAs between PanINs and PDACs. Yellow, PanINs; green, PDACs. Horizontal line indicates median, whiskers indicate minimum and maximum, and box indicates quartiles. $n = 12$ independent PanIN lesions, $n = 8$ PDACs. **b**, Number of SNVs and indels in PanINs and PDACs. **c**, Number of CNAs in PanINs and PDACs (hisens results only).



Extended Data Fig. 2 | Allelic CNAs across all patient samples. CNAs were inferred using the FACETS algorithm (Supplementary Table 3, FACETS. purity variants shown). Blue, putative losses; red, putative gains.

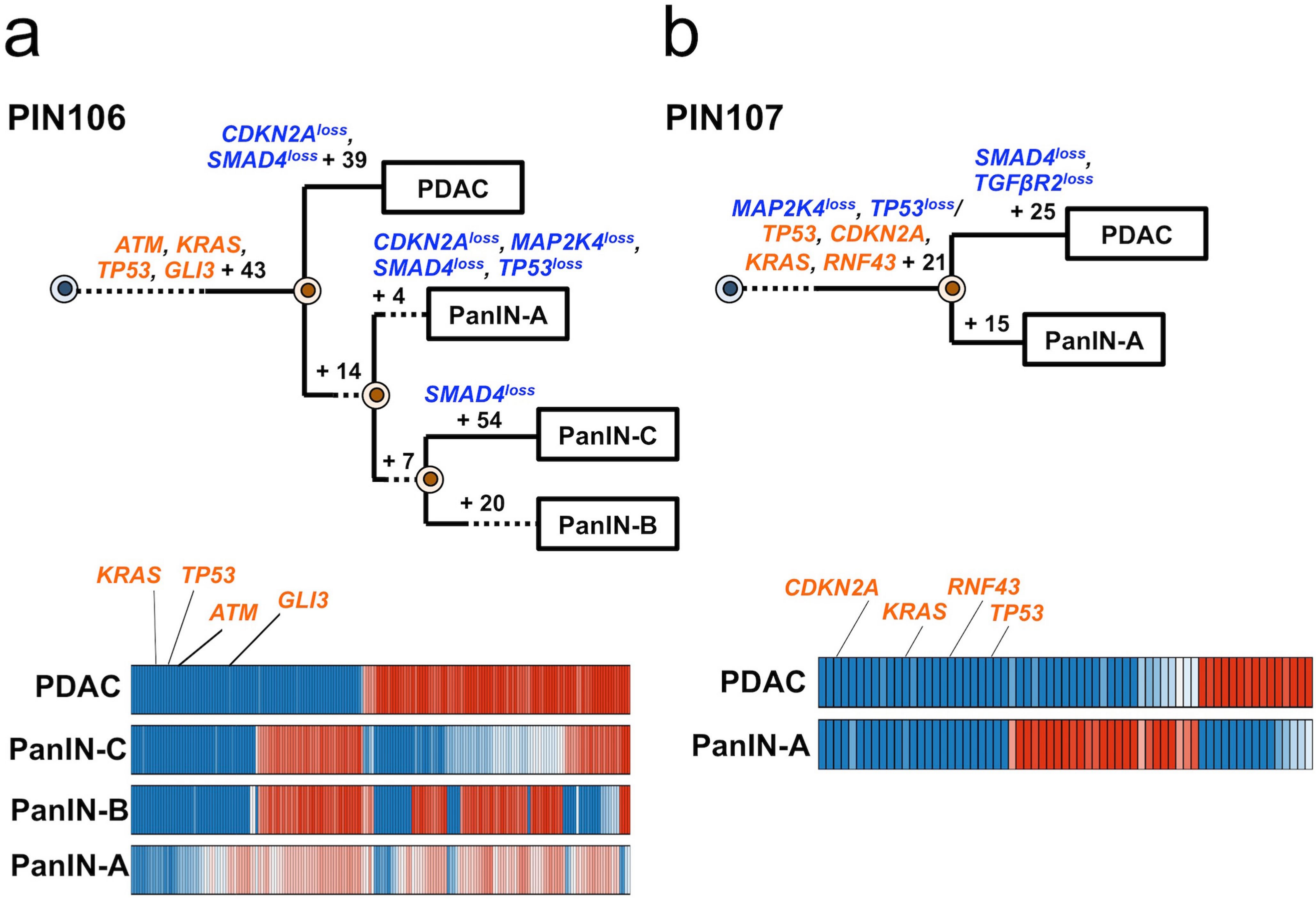

a
PIN102
MYC^gain, SMAD4^loss,
KRAS, CR1 + 128
PDAC
KRAS, SMARCB1 + 32
PanIN-A
KRAS
CR1
SMARCB1
PDAC
PanIN-A
b
PIN105
CDKN2A^loss, MAP2K4^loss, TP53^loss/
TP53, SMAD4^loss/SMAD4, KRAS,
ACVR1B + 76
PDAC
KRAS, CHD4, KMT2D,
ST6GAL2, U2AF1 + 46
PanIN-A
KRAS
ACVR1B
TP53
SMAD4
KMT2D
CHD4
U2AF1
ST6GAL2
PDAC
PanIN-A

Extended Data Fig. 3 | Phylogenetics of PanINs and the matched primary tumour for patients PIN102 and PIN105. See Supplementary Table 1 for sample identities. Gene names in orange text are SNVs or indels, in blue are copy-number losses, and in red are copy-number gains affecting putative driver genes. The sequencing data for each driver gene variant was manually reviewed to verify phylogenetic position. For each phylogeny, the number of acquired mutations is in black. The branch lengths are proportional to the number of SNVs and indels. The dashed line indicates the branch from the germline to the PDAC and PanIN-A. For the Bayesian heat maps, samples are indicated on each row while variants are represented by each column. The colour of each tile indicates the probability that the variant is present or absent in the corresponding sample. Dark blue, >99.9% probability of being present; dark red, >99.9% probability of being absent. Light blue and red indicate lower probabilities; white tiles indicate approximately 50% probability. **a**, Phylogenetic tree and Bayesian heat map with each variant for PIN102. **b**, Phylogenetic tree and Bayesian heat map with each variant for PIN105.

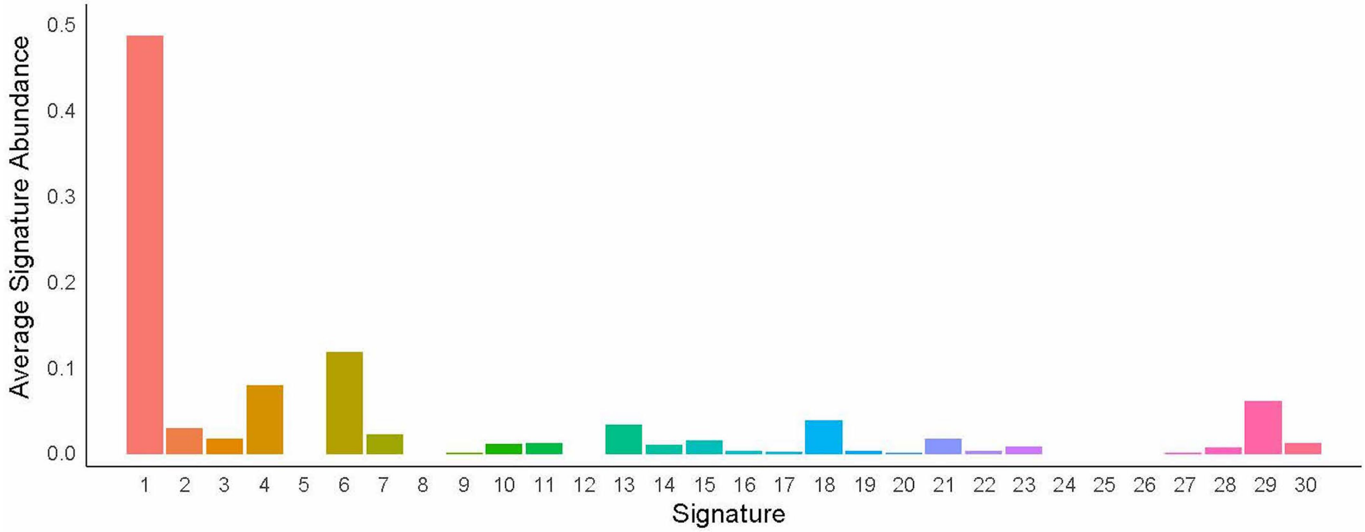

a
PIN101
ACVR1Bloss,
MAP2K4loss,
TP53loss,
SMAD4loss,
KRAS + 13
CDKN2Aloss, MYCgain,
CDKN2A, TP53 + 26
PDAC
SDK2 + 110
PanIN-A
KRAS
TP53
SDK2
CDKN2A
PDAC
PanIN-A
b
PIN103
CDKN2Aloss,
SMAD4loss,
MAP2K4loss,
TP53loss/
TP53 + 24
TGFβR2loss/TGFβR2,
ERBB2, KRAS + 29
PDAC
+ 18
PanIN-A
TP53
TGFβR2
KRAS
ERBB2
PDAC
PanIN-A
c
PIN104
SMAD4loss,
MAP2K4loss,
TP53loss/TP53,
KRAS + 49
SMAD4loss,
GATA6 + 53
PDAC
+22
PanIN-B
+9
PREX2 + 23
PanIN-A
KRAS
TP53
GATA6
PREX2
PDAC
PanIN-B
PanIN-A
d
PIN108
ACVR1Bloss, CDKN2Aloss,
MAP2K4loss, SMAD4loss,
TGFβR2loss, TP53loss,
CNTN5 + 25
PDAC
CDKN2A, KRAS
TP53 + 36
KRASgain,
SMAD4loss + 20
PanIN-A
MYCgain
+ 10
ROBO1,
USP28 + 37
PanIN-B
KRAS
CDKN2A
TP53
CNTN5
USP28
ROBO1
PDAC
PanIN-A
PanIN-B

Extended Data Fig. 4 | See next page for caption.

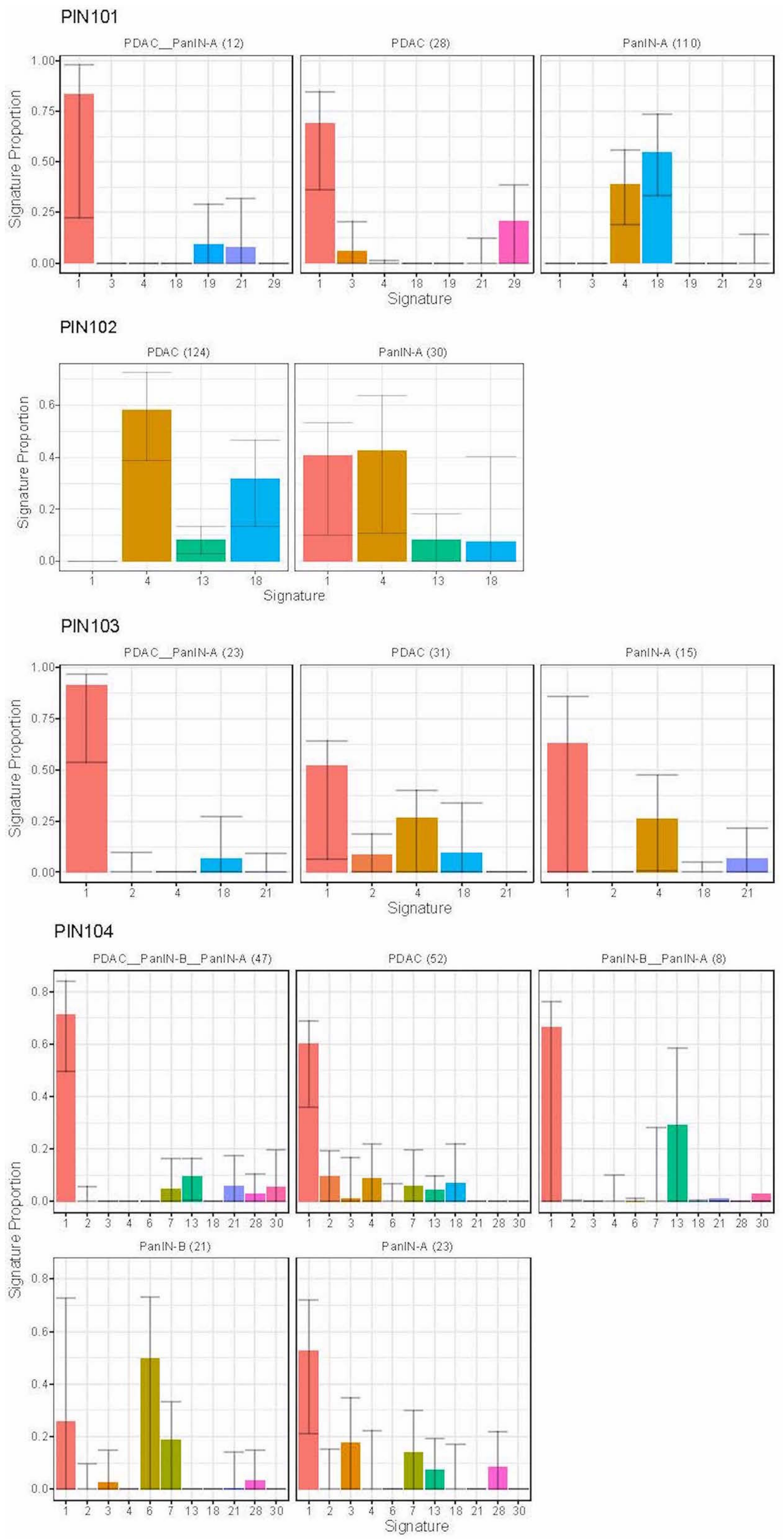
Extended Data Fig. 4 | Phylogenetics of PanINs and matched primary tumours for patients PIN101, PIN103, PIN104, and PIN108. See Supplementary Table 1 for sample identities. Gene names in orange text are SNVs or indels, in blue are copy-number losses, and in red are copy-number gains affecting putative driver genes. The sequencing data for each driver gene variant were manually reviewed to verify phylogenetic position. For each phylogenetic tree, the numbers of acquired mutations are in black. The branch lengths are proportional to the number of SNVs or indels. The dashed lines indicate branches that have been extended to accommodate gene annotation and variant numbers. For the Bayesian heatmaps, samples are indicated on each row and variants are represented by each column. The colour of each tile indicates the probability that the variant is present or absent in the corresponding sample. Dark blue, >99.9% probability of being present; dark red, >99.9% probability of being absent. Light blue and red indicate lower probabilities; white tiles indicate approximately 50% probability. **a**, PIN101. In manual review of the sequencing data, a read supporting the presence of the KRAS(G12D) variant was detected in both the PDAC and PanIN-A samples and was thus moved to the trunk of the phylogeny despite the overall low coverage of *KRAS* in PanIN-A. **b**, PIN103. **c**, PIN104. The node leading from the first MRCA to the second MRCA has a confidence value of >99%. **d**, PIN108. The node leading from the first MRCA to the second MRCA has a confidence value of >99%.


a
PIN106
CDKN2A loss, SMAD4 loss + 39
PDAC
ATM, KRAS, TP53, GLI3 + 43
CDKN2A loss, MAP2K4 loss, SMAD4 loss, TP53 loss
+ 4
PanIN-A
+ 14
SMAD4 loss
+ 54
PanIN-C
+ 7
+ 20
PanIN-B
KRAS
TP53
ATM
GLI3
PDAC
PanIN-C
PanIN-B
PanIN-A
b
PIN107
SMAD4 loss, TGFβR2 loss
+ 25
MAP2K4 loss, TP53 loss/ TP53, CDKN2A, KRAS, RNF43 + 21
PDAC
+ 15
PanIN-A
CDKN2A
KRAS
RNF43
TP53
PDAC
PanIN-A

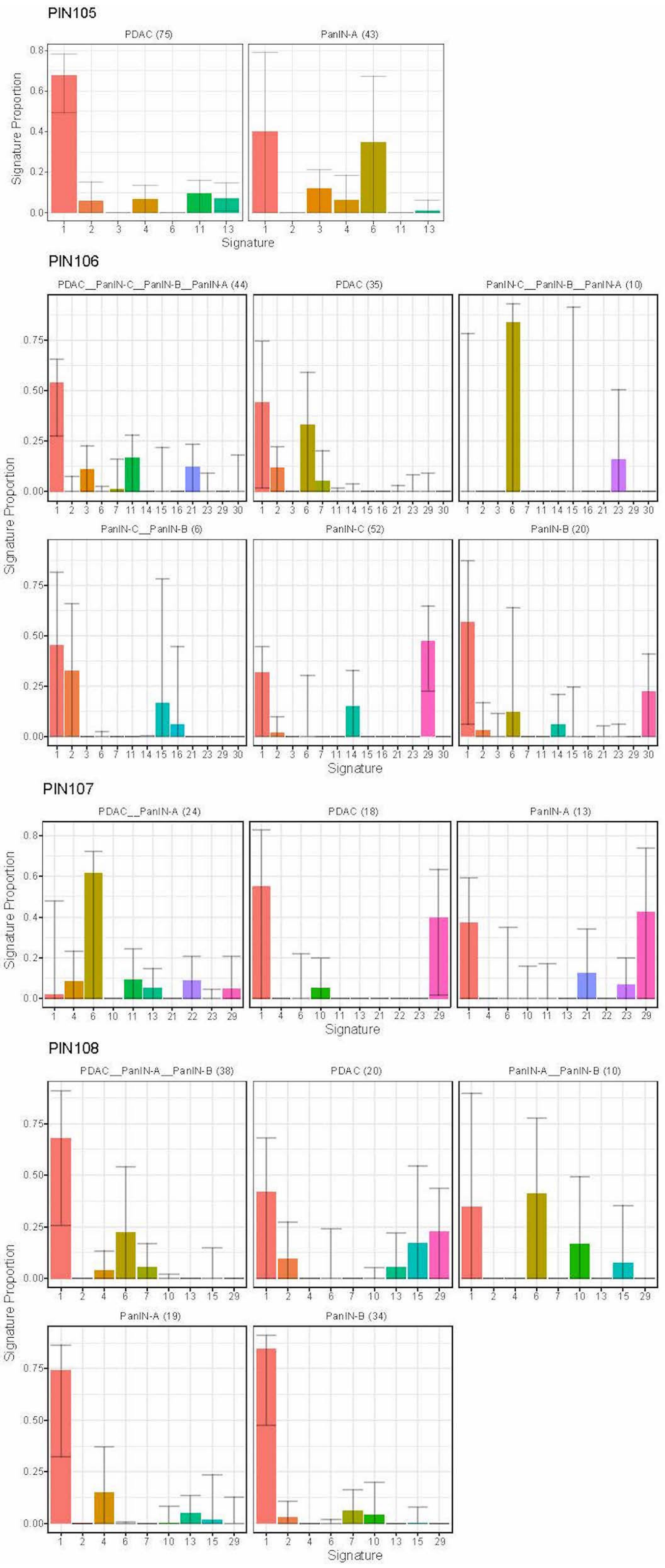
Extended Data Fig. 5 | Phylogenetics of PanINs and matched primary tumours for patients PIN106 and PIN107. See Supplementary Table 1 for sample identities. Gene names in orange text are SNVs or indels, in blue are copy-number losses, and in red are copy-number gains affecting putative driver genes. The sequencing data for each driver gene variant were manually reviewed to verify phylogenetic position. For each phylogeny, the numbers of acquired mutations are in black. The branch lengths are proportional to the number of SNVs or indels. The dashed lines indicate branches that have been extended to accommodate gene annotation and variant numbers. For each Bayesian heat map, samples are indicated on each row while variants are represented by each column. The colour of each tile indicates the probability that the variant is present or absent in the corresponding sample. Dark blue, >99.9% probability of being present; dark red, >99.9% probability of being absent. Light blue and red indicate lower probabilities; white tiles indicate approximately 50% probability. **a**, PIN106. The node leading from the first MRCA to the second MRCA has a confidence value of >99% and the node leading from the second MRCA to the third MRCA has a confidence value of 82%. **b**, PIN107.



Extended Data Fig. 6 | Average signature abundance across samples. Signature numbers 1–30 from Alexandrov et al.[43] are shown on the *x*-axis with signature abundance averaged across phylogenetic branches shown on the *y*-axis. Each histogram is coloured by signature identity.



Extended Data Fig. 7 | The proportion of mutational signatures from Alexandrov et al.[43] estimated in PIN101–PIN104. Signatures are shown on the *x*-axis, with the proportion of each signature shown on the *y*-axis. Each bar is coloured by signature identity. The text on the top of each panel denotes the corresponding phylogenetic branch and the number of mutations acquired along it in parentheses. Error bars depict 90% CIs in the signature proportion estimated by 100 iterations of bootstrap resampling.



Extended Data Fig. 8 | The proportion of mutational signatures from Alexandrov et al.[43] estimated in PIN105–PIN108. Signatures are shown on the *x*-axis, with the proportion of each signature shown on the *y*-axis. Each bar is coloured by signature identity. The text on the top of each panel denotes the corresponding phylogenetic branch and the number of mutations acquired along it in parentheses. Error bars depict 90% CIs in the signature proportion estimated by 100 iterations of bootstrap resampling.

LETTER

https://doi.org/10.1038/s41586-018-0472-9

Autogenous and efficient acceleration of energetic ions upstream of Earth's bow shock

D. L. Turner[1]*, L. B. Wilson III[2], T. Z. Liu[3], I. J. Cohen[4], S. J. Schwartz[5], A. Osmane[6,7], J. F. Fennell[1], J. H. Clemmons[1], J. B. Blake[1], J. Westlake[4], B. H. Mauk[4], A. N. Jaynes[8], T. Leonard[9], D. N. Baker[9], R. J. Strangeway[3], C. T. Russell[3], D. J. Gershman[2], L. Avanov[2], B. L. Giles[2], R. B. Torbert[10,11], J. Broll[11,12], R. G. Gomez[1,1], S. A. Fuselier[11,12] & J. L. Burch[11]

Earth and its magnetosphere are immersed in the supersonic flow of the solar-wind plasma that fills interplanetary space. As the solar wind slows and deflects to flow around Earth, or any other obstacle, a 'bow shock' forms within the flow. Under almost all solar-wind conditions, planetary bow shocks such as Earth's are collisionless, supercritical shocks, meaning that they reflect and accelerate a fraction of the incident solar-wind ions as an energy dissipation mechanism[1,2], which results in the formation of a region called the ion foreshock[3]. In the foreshock, large-scale, transient phenomena can develop, such as 'hot flow anomalies'[4–9], which are concentrations of shock-reflected, suprathermal ions that are channelled and accumulated along certain structures in the upstream magnetic field. Hot flow anomalies evolve explosively, often resulting in the formation of new shocks along their upstream edges[5,10], and potentially contribute to particle acceleration[11–13], but there have hitherto been no observations to constrain this acceleration or to confirm the underlying mechanism. Here we report observations of a hot flow anomaly accelerating solar-wind ions from roughly 1–10 kiloelectronvolts up to almost 1,000 kiloelectronvolts. The acceleration mechanism depends on the mass and charge state of the ions and is consistent with first-order Fermi acceleration[14,15]. The acceleration that we observe results from only the interaction of Earth's bow shock with the solar wind, but produces a much, much larger number of energetic particles compared to what would typically be produced in the foreshock from acceleration at the bow shock. Such autogenous and efficient acceleration at quasi-parallel bow shocks (the normal direction of which are within about 45 degrees of the interplanetary magnetic field direction) provides a potential solution to Fermi's 'injection problem', which requires an as-yet-unexplained seed population of energetic particles, and implies that foreshock transients may be important in the generation of cosmic rays at astrophysical shocks throughout the cosmos.

Since the start of its first science phase in September 2015, NASA's Magnetospheric Multiscale mission (MMS)[16] has provided an unprecedented set of observatories with which transient foreshock phenomena can be studied. MMS involves four identically instrumented spacecraft (MMS-1–4) that are held in a tight, tetrahedron formation with inter-satellite separations ranging throughout the mission from less than 10 km to around 100 km. The four-point configuration enables multipoint analysis and the disambiguation of spatiotemporal features in observables that include plasma, fields and waves, energetic particles and ion composition at resolutions not previously available for in situ space plasma experiments. Here, we report on MMS observations of a hot flow anomaly (HFA) from 28 December 2015 that was associated with an unusual observation of energetic ions up to several hundred kiloelectronvolts in the solar wind and magnetosheath (the region of shocked solar-wind plasma flowing around Earth's magnetosphere downstream of the bow shock). Using the multipoint nature of the mission and its comprehensive suite of instrumentation, we are able to examine this event in detail to determine whether and how the accelerated ions are related to the HFA and its interaction with the bow shock.

HFAs are characterized by superheated, tenuous, low-field-strength core regions that exhibit strong deflections of the plasma velocity[17]. As the structure responsible for their formation convects away from the Sun with the solar wind through the foreshock, HFA cores intensify and, in an attempt to maintain pressure balance, expand into the surrounding solar-wind plasma, forming compression regions on either side and often resulting in the formation of fast magnetosonic shocks at their upstream edges. In Fig. 1 we show 50 s of MMS burst magnetic field and ion and electron plasma data encompassing a transit of the spacecraft through a well-developed HFA. MMS first encountered the downstream edge of the HFA (with respect to the flow of the solar wind away from the Sun) around 05:26:53 UT, where the enhanced density and magnetic-field strength indicated a region of compressed solar-wind plasma. The spacecraft passed into the core of the HFA from about 05:26:58 UT to 05:27:25 UT, observing plasma that was highly tenuous (density of less than 1 cm^{-3}), had a low field strength ($|\boldsymbol{B}| < 2$ nT) and was strongly deflected ($V_z > 300$ km s^{-1} and $V_x > -200$ km s^{-1}; the velocity $\boldsymbol{V}$ was towards the Sun for several seconds). MMS next encountered shocked plasma and strong magnetic-field variations just before the spacecraft passed through a fast magnetosonic shock at the upstream edge of the HFA at around 05:27:29 UT.

Unusually high intensities of energetic ions were observed by the Energetic Ion Spectrometer (EIS) and Fly's Eye Energetic Particle Spectrometer (FEEPS) instruments on MMS during and for several minutes following the encounter with the HFA. Figure 2 reveals an enhancement in the number of protons and of helium and heavier (carbon–nitrogen–oxygen, CNO) ions with energies of more than 50 keV starting at around 05:27 UT, which coincides with the initial encounter with the HFA, and persisting until about 05:33 UT, after MMS exited the HFA and transited the bow shock into the magnetosheath. The peak intensities and the upper energy thresholds in the energy distributions of these ions increased exponentially over time, and the rate of increase and the upper energy limits were mass and/or charge dependent: the heaviest ions observed (CNO^{n+}) exhibited a faster rate of energy increase and a higher upper energy limit (up to about 800 keV) than did the helium ions (He^{n+}, energy limit up to about 400 keV) and protons (H^{+}, energy limit less than 200 keV). From the highest energy channels (roughly 10–40 keV) of the Hot Plasma Composition Analyzers (HPCAs) on MMS, which can distinguish the charge state of helium and heavier ions (see Extended Data Fig. 1), we determined that the helium ions observed were clearly α particles (He^{2+}). In addition, there were no observable levels of O^{+},

---

[1]Space Sciences Department, The Aerospace Corporation, El Segundo, CA, USA. [2]NASA Goddard Space Flight Center, Greenbelt, MD, USA. [3]Department of Earth, Planetary, and Space Science, University of California, Los Angeles, CA, USA. [4]Applied Physics Laboratory, Laurel, MD, USA. [5]Imperial College London, London, UK. [6]School of Electrical Engineering, Aalto University, Espoo, Finland. [7]Rudolf Peierls Centre of Theoretical Physics, University of Oxford, Oxford, UK. [8]Department of Physics and Astronomy, University of Iowa, Iowa City, IA, USA. [9]Laboratory for Atmospheric and Space Physics, University of Colorado, Boulder, CO, USA. [10]Institute For the Study of Earth, Oceans, and Space, University of New Hampshire, Durham, NH, USA. [11]Southwest Research Institute, San Antonio, TX, USA. [12]Departoment of Physics and Astronomy, University of Texas at San Antonio, San Antonio, TX, USA. *e-mail: drew.l.turner@aero.org

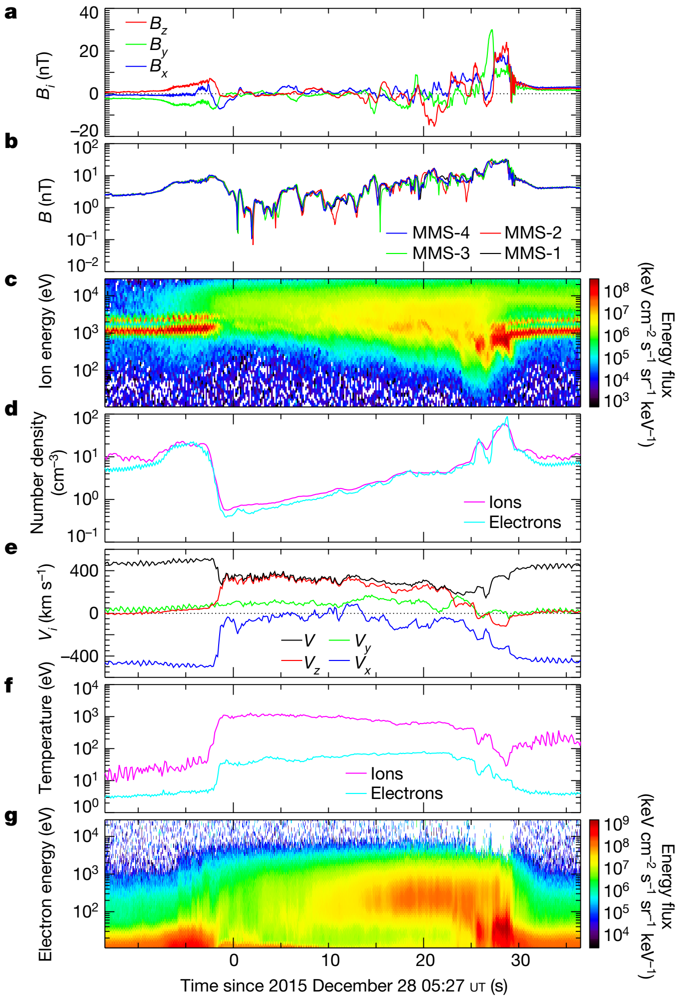


Fig. 1 | Overview of the MMS burst data during the HFA. a–g, Quantities shown (all from MMS-1 unless otherwise specified) are the components of the magnetic-field vector $\boldsymbol{B} = (B_x, B_y, B_z)$ (**a**) and the magnetic-field strength $B = |\boldsymbol{B}|$ from the flux-gate magnetometers on each of the four spacecraft (**b**); the omni-directional, averaged energy distributions of the ions from the fast plasma investigation dual ion spectrometer (FPI/DIS) (**c**); the ion and electron densities (**d**); the components of the velocity vector of the ions (V_x, V_y, V_z) and its magnitude $V = |\boldsymbol{V}|$ (**e**); the ion and electron temperatures (**f**); and the omni-directional, averaged energy distributions of the electrons from the FPI dual electron spectrometer (FPI/DES) (**g**). With the geometry and multipoint observations of the HFA, its size, expansion rate and motion along the bow shock can be calculated. Using the rotation of the IMF before and after the HFA, we estimate the normal direction of the IMF discontinuity that is responsible for the formation of the HFA to be $\boldsymbol{n} = (0.70, -0.44, -0.57)$ in the geocentric solar ecliptic coordinate system. We calculated the normal direction of the upstream shock of the HFA using co-planarity and multi-spacecraft analysis (Methods), yielding $\boldsymbol{n}_{\mathrm{sh}} = (0.85, -0.16, -0.53)$ and $\boldsymbol{n}_{\mathrm{sh}} = (0.84, -0.44, -0.31)$, respectively, which are generally consistent. In the solar-wind frame, the HFA shock was expanding upstream (towards the Sun) at about 360 km s^{-1}. When MMS observed the HFA, it was expanding at a rate of 108 km s^{-1} along the IMF discontinuity normal and was approximately 16,000 km (2.5 Earth radii) in size from edge to edge along the track taken by MMS through the HFA. The geometry of the HFA and the values that we have calculated are critical for enabling comparisons between observed particle acceleration signatures and theory.

which indicates that the heaviest energetic ions observed were also at a high charge state (such as O^{6+} or C^{6+}, the predominant heavy ions and charge states in the CNO branch in the solar wind). Both of those observations are consistent with these accelerated ions being of solar-wind origin.

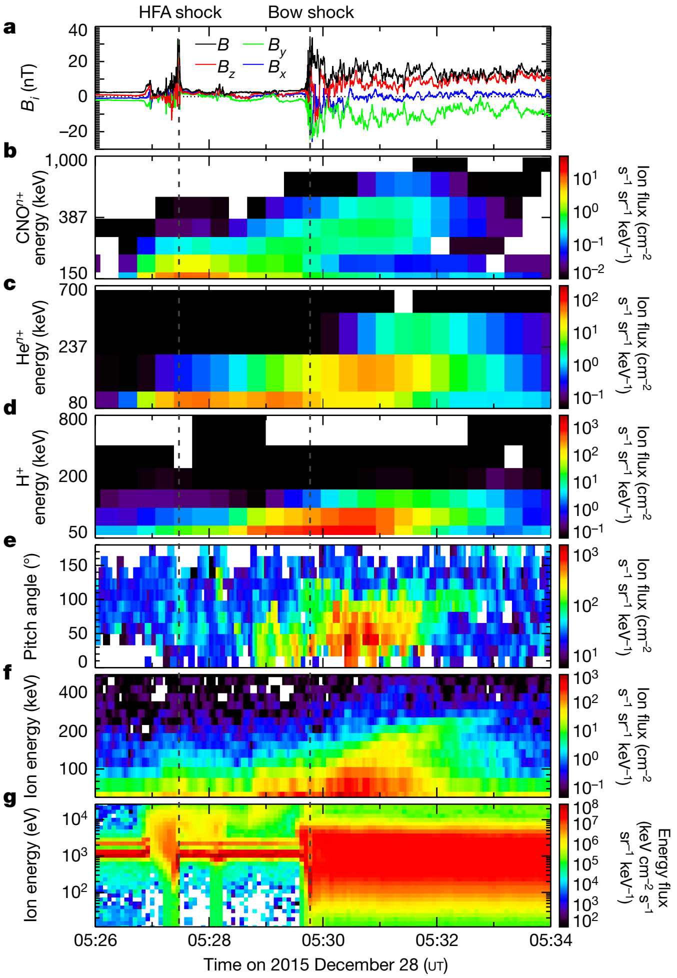


Fig. 2 | MMS observations of energetic ions and context data during and after the encounter with the HFA. b–d, Heavy-ion (CNO^{n+}; **b**), helium (He^{n+}; **c**) and proton (H$^+$; **d**) data from the energetic particle detector EIS (EPD/EIS). The colour scales show the differential fluxes and the energies shown (on the y axes) are the total energy per incident ion. **e**, Pitch angle distributions for 70–105-keV ions from the EPD/FEEPS instruments. For **b**–**e**, the data were compiled from counts on the EIS or FEEPS instruments on all four spacecraft. **f**, **g**, Total ion energy distributions from all four FEEPS instruments (**f**) and the DIS instrument on MMS-3 (**g**). **a**, Magnetic-field vector components and strength from MMS-3, for context. The grey, dashed, vertical lines mark the HFA shock and the bow shock, as labelled. The energetic ions were first observed within the HFA around 05:27 UT and then continuously after the HFA through the ion foreshock and into the magnetosheath until around 05:33 UT.

From the pitch-angle distributions from FEEPS, we determined that the energetic ions were observed streaming along magnetic field lines between about 05:29 UT and 05:32 UT, when the interplanetary magnetic field (IMF) was generally steady, with an average of (1.1, −2.6, 1.6) ± (0.4, 0.4, 0.5) nT in the solar wind and (−1.1, −11.1, 7.7) ± (2.7, 3.6, 2.6) nT in the magnetosheath. HPCA also observed perpendicularly heated, field-aligned beams of He^{2+} and H$^+$ ions with energies of more than 10 keV that are consistent with the energetic ion distributions observed by EIS and FEEPS. The pitch-angle distributions and field orientations are consistent with ions streaming along field lines away from the location of the HFA, as illustrated in Fig. 3. In addition, the magnetic-field orientation in the magnetosheath was nearly perpendicular to the magnetopause-boundary normal on the side of the system adjacent to MMS, providing additional evidence that the accelerated ions (H$^+$, He^{2+} and CNO^{n+}) were not of magnetospheric origin. Furthermore, magnetospheric escape is inconsistent with the continuous, high intensities that originated in and continued after the HFA and the limited extent of these ions within the magnetosheath.

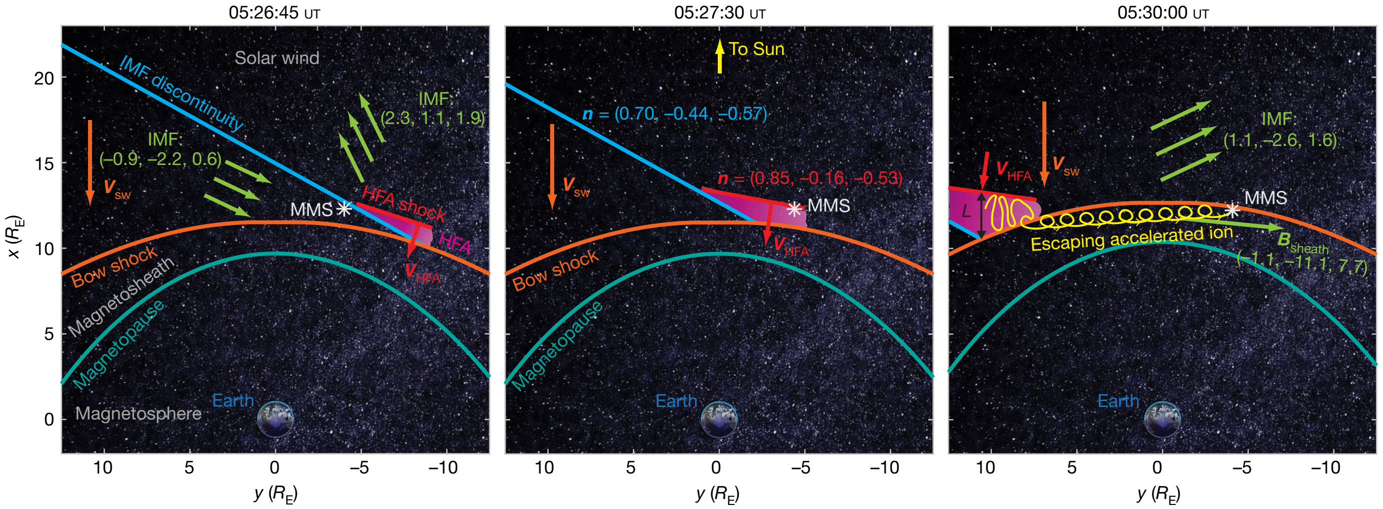


Fig. 3 | Conceptual scenario for the HFA observed by MMS on 28 December 2015. Earth is shown at the origin of three two-dimensional cross-sections (approximately in Earth's equatorial plane; $R_{\mathrm E}$ is the radius of Earth), which capture the geometry of the system at 05:26:45 UT, 05:27:30 UT and 05:30:00 UT (left to right). The position of MMS is marked by the white asterisk. Slices of Earth's magnetopause, Earth's bow shock, the IMF discontinuity and the HFA shock are shown in turquoise, orange, blue and red, respectively. Velocity vectors for the solar wind ($\boldsymbol{V}_{\mathrm{sw}}$) and the HFA shock ($\boldsymbol{V}_{\mathrm{HFA}}$) are also shown. IMF geometries and vectors are shown in green, and the normal vectors of the IMF discontinuity ($\boldsymbol{n} = (0.70, -0.44, -0.57)$) and the HFA shock ($\boldsymbol{n} = (0.85, -0.16, -0.53)$) are also labelled. The core of the HFA is shown in magenta. As the IMF discontinuity and HFA convect with the solar wind, they sweep from right to left across the bow shock, as shown here. The HFA grows in time as it sweeps up and concentrates more and more suprathermal ions in the quasi-parallel foreshock that forms on the upstream side (towards the Sun) of the IMF discontinuity. In the solar-wind rest frame, Earth's bow shock is moving towards the Sun at V_{sw} (orange velocity vector) and is perpetually converging on and overtaking the IMF discontinuity and the HFA with its shock (moving at $V_{\mathrm{HFA}} < V_{\mathrm{sw}}$ and shown as the red velocity vector). This situation enabled continuous first-order Fermi acceleration to occur during the life of the HFA. In this case, we were afforded a period of fortuitous IMF orientation after MMS observed the HFA: ions that were accelerated within the Fermi acceleration trap between the HFA shock and the bow shock escaped the trap and were then observed by MMS streaming away from the HFA, allowing for direct, observationally constrained comparison between the observations and theory. The yellow spiral trace in the 05:30:00 UT snapshot shows an ion escaping the trap and streaming towards MMS, which is consistent with the observed FEEPS ion and high-energy HPCA He^{2+} and H^{+} angular distributions. L is the length scale between the HFA shock and the bow shock, and the average field in the magnetosheath observed by MMS is shown as $\boldsymbol{B}_{\mathrm{sheath}}$. Image of Earth from NASA's From Earth to the Solar System collection (image credit: Data-AVHRR, NDVI, Seawifs, MODIS, NCEP, DMSP and Sky2000 star catalogue; AVHRR and Seawifs texture-Reto Stockli; Visualization-Marit Jentoft-Nilsen).

Using these observations, we test the hypothesis that these energetic ions were accelerated in—and subsequently observed after escaping from—a Fermi-type acceleration 'trap' between two converging magnetic mirrors, that is, the HFA shock or sheath and Earth's bow shock or magnetosheath. Following Fermi's work on particle acceleration in astrophysical plasmas, a relationship can be derived that relates the maximum energy, mass and charge state of an ion for any particular Fermi acceleration trap (equation (2) in Methods). The energy thresholds (Fig. 4, dashed vertical lines) derived from this theoretical relationship, calculated for protons, α particles and heavier CNO ions, agree well with the energy spacing between the observed peaks in the energetic ion spectra from MMS, to within the energy resolution of the EIS instruments.

We elaborate on this comparison to theory using an independent and more advanced model for ions interacting with an HFA shock or sheath of finite thickness. In this model (Methods), we consider the gyroradius, gyrophase and pitch angle of an ion as it intersects the HFA shock or sheath region, so that finite-gyroradius effects are included whereby only a fraction of ions that intersect this region are reflected and accelerated by it and remain within the acceleration trap. With this model, the energy spectrum of accelerated ions that escape the trap can be estimated for each species, because the distributions also depend on the mass and charge of the ions. Solutions to this model are shown in Fig. 4 (dash-dotted curves) and are in excellent agreement with the observed spectra of the accelerated ions considering the energy resolution and one-count levels of the EIS instruments. FEEPS data, which do not distinguish different ion species, captured the multiple 'knees' in the energetic ion distribution, also in excellent agreement with the theoretical spectra for the three ion species.

A third, independent, theoretical test estimates the length scale of a Fermi acceleration trap using the average rate of energy increase and momentum of the accelerated particles and the convergence speed of the trap (Methods). With the observed values of these quantities, we can estimate the length scale between the HFA shock or sheath region and Earth's bow shock or magnetosheath. If Fermi acceleration between the two was the mechanism of acceleration, then the length scale that we calculate should be comparable to that observed by MMS within the HFA. We find a length scale of 3.2 Earth radii, which is remarkably consistent with that observed by MMS (2.5 Earth radii), especially considering any growth of the HFA over time.

On the basis of the observational evidence and excellent agreement with theory, we conclude that the energetic ions observed by MMS during and after the HFA were accelerated via a first-order Fermi acceleration process between the converging HFA shock or sheath and Earth's bow shock or magnetosheath. Figure 3 illustrates the conceptual scenario in three cross-sectional snapshots for the observed geometry of the system. State-of-the-art global hybrid models that are capable of simulating HFAs at planetary bow shocks may not capture such acceleration owing to limitations in their system sizes or spatial and temporal resolution at sub-ion scales, particularly at shock fronts, in turbulent sheaths and within simulated HFAs. Our observations link intense bursts of energetic ions in the foreshock and magnetosheath to HFAs, which occur frequently (several per day) upstream of Earth's bow shock, particularly for fast solar wind with radial IMF[9,18,19]. Because HFAs are localized structures, it is possible that spacecraft can observe the particles accelerated by an HFA remotely, without ever observing the HFA itself. Shocks are important sites of particle acceleration[20–24], and our observational confirmation of how ions are accelerated in

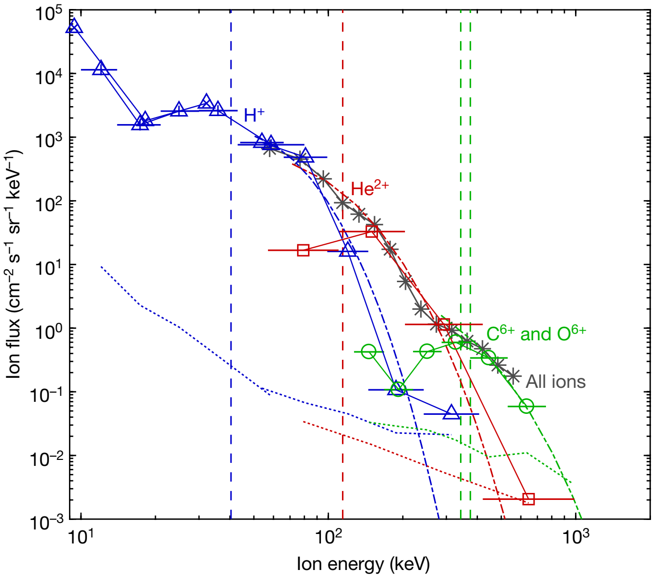


Fig. 4 | Comparing observations and theory. Energy distributions of ions with energies of tens to hundreds of kiloelectronvolts determined from the HPCA, EIS and FEEPS instruments on all four MMS spacecraft for the period 05:30:00–05:31:25 UT are shown as solid lines and markers. Data for heavy (CNO branch) and helium ions are shown as green circles and red squares, respectively. Proton data from HPCA are shown as blue triangles around blue crosses; proton data from the EIS energy-by-time-of-flight ($E \times$ TOF) technique are shown as open blue triangles; and proton data from the EIS pulse-height-by-time-of-flight (PH $\times$ TOF) technique are shown as blue triangles around blue plus signs. On all EIS data points, the solid horizontal lines show the energy range of each channel. The dotted lines show one-count background levels for the EIS data. EIS PH $\times$ TOF data have been cross-calibrated to HPCA and $E \times$ TOF data where there is overlap between those instruments. FEEPS data, which do not distinguish between different ion species, are shown as grey lines and asterisks. FEEPS data have had a constant cross-calibration factor of two applied to all energy channels. The dashed vertical lines represent $E_{\max}$ from equation (2) in Methods (the maximum energy in a Fermi acceleration trap of size L) for protons (H^+, blue), α particles (He^{2+}, red) and heavy ions (C^{6+} and O^{6+}, green) using a constant in equation (2) in Methods calculated for 40-keV protons (with mass $m = 1$ and charge $q = 1$). The dash-dotted, exponentially decaying lines show numerical solutions for the high-energy tails of the distributions of protons (blue), α particles (red) and O^{6+} ions (green) accelerated between the observed HFA and Earth's bow shock, using the Fermi acceleration model derived in Methods.

HFAs has important implications for particle acceleration at supercritical, collisionless shocks in other astrophysical plasmas.

Unlike the traditional understanding of diffusive shock acceleration, in which particles are scattered randomly by kinetic-scale waves propagating on either side of a single shock, the more efficient form of Fermi acceleration that we have identified here involves two large-scale shocks that converge continuously on each other throughout the life of the HFA. The supercritical bow shock interacting with the inhomogeneous solar wind (which is responsible for the formation of the bow shock) results in the formation of the HFA, and HFAs can also form independently of IMF discontinuities[25,26]. Thus, the Fermi acceleration of ions—as observed here—occurs autogenously within the ion foreshock, independently of any interaction with an external system (such as an interplanetary shock[27]). This result provides important insights into and suggests new factors to consider concerning the 'injection problem' of shock acceleration[28]. Furthermore, the newly identified acceleration mechanism may be even more effective in systems with more uniformly planar shocks (such as interplanetary shocks, astrospherical shocks and supernova shocks), in which HFAs can interact with the parent shock for much longer time scales and any curvature of the parent shock is perpetually much greater than the gyroradii of the accelerated ions. The discovery that first-order Fermi acceleration between two shocks occurs autogenously upstream of a supercritical, collisionless shock implies that quasi-parallel foreshock regions (where the interplanetary magnetic field and the normal direction of the bow shock are within about 45° of each other) and the foreshock transients that form within them, such as HFAs, may be important in particle acceleration and the generation of cosmic rays at other astrophysical shocks throughout the cosmos.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0472-9.

Received: 28 March; Accepted: 6 July 2018;
Published online 12 September 2018.

1. Wilson, L. B. III et al. Quantified energy dissipation rates in the terrestrial bow shock: 1. Analysis techniques and methodology. *J. Geophys. Res.* **119**, 6455–6474 (2014).
2. Wilson, L. B. III et al. Quantified energy dissipation rates in the terrestrial bow shock: 2. Waves and dissipation. *J. Geophys. Res.* **119**, 6475–6495 (2014).
3. Eastwood, J. P. et al. The foreshock. *Space Sci. Rev.* **118**, 41–94 (2005).
4. Schwartz, S. J. et al. An active current sheet in the solar wind. *Nature* **318**, 269–271 (1985).
5. Thomsen, M. F. et al. On the origin of hot diamagnetic cavities near the Earth's bow shock. *J. Geophys. Res.* **93**, 11311–11325 (1988).
6. Paschmann, G. et al. Three-dimensional plasma structures with anomalous flow directions near the Earth's bow shock. *J. Geophys. Res.* **93**, 11279–11294 (1988).
7. Schwartz, S. J. et al. Conditions for the formation of hot flow anomalies at Earth's bow shock. *J. Geophys. Res.* **105**, 12639–12650 (2000).
8. Omidi, N. & Sibeck, D. G. Formation of hot flow anomalies and solitary shocks. *J. Geophys. Res.* **112**, A01203 (2007).
9. Wang, S., Zong, Q. & Zhang, H. Hot flow anomaly formation and evolution: cluster observations. *J. Geophys. Res.* **118**, 4360–4380 (2013).
10. Fuselier, S. A. et al. Fast shocks at the edges of hot diamagnetic cavities upstream from the Earth's bow shock. *J. Geophys. Res.* **92**, 3187–3194 (1987).
11. Wilson, L. B. III et al. Relativistic electrons produced by foreshock disturbances observed upstream of Earth's bow shock. *Phys. Rev. Lett.* **117**, 215101 (2016).
12. Liu, T. Z. et al. Statistical study of particle acceleration in the core of foreshock transients. *J. Geophys. Res.* **122**, 7197–7208 (2017).
13. Giacalone, J. & Burgess, D. Interaction between inclined current sheets and the heliospheric termination shock. *Geophys. Res. Lett.* **37**, L19104 (2010).
14. Fermi, E. On the origin of cosmic radiation. *Phys. Rev.* **75**, 1169–1174 (1949).
15. Liu, T. Z. et al. Fermi acceleration of electrons inside foreshock transient cores. *J. Geophys. Res.* **122**, 9248–9263 (2017).
16. Burch, J. L. et al. Magnetospheric Multiscale overview and science objectives. *Space Sci. Rev.* **199**, 5–21 (2016).
17. Liu, T. Z., Turner, D. L., Angelopoulos, V. & Omidi, N. Multipoint observations of the structure and evolution of foreshock bubbles and their relation to hot flow anomalies. *J. Geophys. Res.* **121**, 5489–5509 (2016).
18. Turner, D. L. et al. First observations of foreshock bubbles upstream of Earth's bow shock: characteristics and comparisons to HFAs. *J. Geophys. Res.* **118**, 1552–1570 (2013).
19. Chu, C. et al. THEMIS satellite observations of hot flow anomalies at Earth's bow shock. *Ann. Geophys.* **35**, 443–451 (2017).
20. Jones, F. C. & Ellison, D. C. The plasma physics of shock acceleration. *Space Sci. Rev.* **58**, 259–346 (1991).
21. Caprioli, D., Pop, A.-R. & Spitkovsky, A. Simulations and theory of ion injection at non-relativistic collisionless shocks. *Astrophys. J.* **798**, L28 (2015).
22. Giacalone, J. Particle acceleration at shocks moving through an irregular magnetic field. *Astrophys. J.* **624**, 765–772 (2005).
23. Park, J., Caprioli, D. & Spitkovsky, A. Simultaneous acceleration of protons and electrons at nonrelativistic quasiparallel collisionless shocks. *Phys. Rev. Lett.* **114**, 085003 (2015).
24. Caprioli, D., Yi, D. T. & Spitkovsky, A. Chemical enhancements in shock-accelerated particles: ab initio simulations. *Phys. Rev. Lett.* **119**, 171101 (2017).
25. Zhang, H. et al. Spontaneous hot flow anomalies at quasi-parallel shocks: 1. Observations. *J. Geophys. Res.* **118**, 3357–3363 (2013).
26. Omidi, N., Zhang, H., Sibeck, D. & Turner, D. Spontaneous hot flow anomalies at quasi-parallel shocks: 2. Hybrid simulations. *J. Geophys. Res.* **118**, 173–180 (2013).
27. Hietala, H., Sandroos, A. & Vainio, R. Particle acceleration in shock-shock interaction: model to data comparison. *Astrophys. J.* **751**, L14 (2012).
28. Balogh, A. & Treumann, R. A. *Physics of Collisionless Shocks* (Springer, New York, 2013).

Acknowledgements This work was supported by NASA contract NNG04EB99C at Southwest Research Institute, a NASA grant (NNX16AQ50G) and research supported by the International Space Science Institute's (ISSI) International Teams programme. We thank all of the MMS team and the SPEDAS software developers for their publicly available data and software products. D.L.T. thanks T. Phan, S.-H. Lee and D. G. Sibeck for discussions and NASA's From Earth to the Solar System collection.

Reviewer information *Nature* thanks H. Zhang and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions D.L.T. performed the data analysis, interpretation and manuscript preparation. L.B.W., T.Z.L., S.J.S. and A.O. contributed to data interpretation, multipoint analysis and development of the theory for comparison to the observations. J.F.F., J.H.C., J.B.B., A.N.J., T.L. and D.N.B. contributed to the development, operation and data processing of the FEEPS energetic particle telescopes, data quality assurance and interpretation of those data. I.J.C., J.W. and B.H.M. contributed to the development, operation and data processing of the EIS instruments, data quality assurance and interpretation of those data. R.J.S. and C.T.R. contributed to the development, operation and data processing of the fluxgate magnetometers and related data quality assurance. D.J.G., L.A. and B.L.G. contributed to the development, operation and data processing of the fast plasma instruments and related data quality assurance. R.B.T. contributed to the development, operation and data processing of the FIELDS instrument suite and related data quality assurance. J.B., R.G.G. and S.A.F. contributed to the development operation and data processing of the HPCA instruments, data quality assurance and interpretation of those data. J.L.B. is the PI of MMS science and contributed with quality assurance and interpretation of the data.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0472-9.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to D.L.T.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Data handling. For this study, MMS data from the following instrument suites were used: FIELDS (electric and magnetic fields and waves)[29], EPD (energetic particle distributions)[30,31], FPI (electron and ion plasma distributions and moments)[32] and HPCA (ion plasma composition, distributions and moments)[33]. All data used for this study were from the standard level-2 data products, which are publicly available at the MMS Science Data Center (https://lasp.colorado.edu/mms/sdc/public).

Multipoint analysis. Much of the analysis of the HFA examined here follows previous work[7], with additional multipoint analysis techniques[34] and analysis techniques for shock and discontinuity normals[35] detailed elsewhere.

First, to determine the normal direction of the IMF discontinuity responsible for the formation of the HFA, cross-product analysis was performed using the $\boldsymbol{B}$-field observations before and after the HFA[35]. These vectors were $\boldsymbol{B}_{\mathrm{pre}}=(-0.89, -2.24, 0.62)$ nT from 05:26:25 UT to 05:26:45 UT and $\boldsymbol{B}_{\mathrm{post}}=(2.25, 1.14, 1.89)$ nT from 05:27:35 UT to 05:27:55 UT, both in geocentric solar ecliptic (GSE) coordinates, yielding a discontinuity (current sheet) normal direction of $\boldsymbol{n}=(0.70, -0.44, -0.57)$ in GSE coordinates. This current-sheet orientation satisfies the HFA formation criterion that the convection electric field is pointed back into the current sheet on at least one side.

For the normal direction of the HFA shock, we compared two methods[35]: shock co-planarity and multi-spacecraft timing analysis. For shock co-planarity, the normal direction of a shock can be determined using the upstream and downstream magnetic-field and velocity vectors:

$$\boldsymbol{n}_{\mathrm{sh}}=\pm\frac{[(\boldsymbol{B}_{\mathrm{up}}-\boldsymbol{B}_{\mathrm{dn}})\times(\boldsymbol{V}_{\mathrm{up}}-\boldsymbol{V}_{\mathrm{dn}})]\times(\boldsymbol{B}_{\mathrm{up}}-\boldsymbol{B}_{\mathrm{dn}})}{|[(\boldsymbol{B}_{\mathrm{up}}-\boldsymbol{B}_{\mathrm{dn}})\times(\boldsymbol{V}_{\mathrm{up}}-\boldsymbol{V}_{\mathrm{dn}})]\times(\boldsymbol{B}_{\mathrm{up}}-\boldsymbol{B}_{\mathrm{dn}})|}$$

with upstream conditions from 05:27:32 UT to 05:27:34 UT and downstream conditions from 05:27:28.0 UT to 05:27:28.8 UT such that $\boldsymbol{B}_{\mathrm{up}}=(2.87, 1.74, 2.31)$ nT, $\boldsymbol{B}_{\mathrm{dn}}=(15.33, 11.56, 21.85)$ nT, $\boldsymbol{V}_{\mathrm{up}}=(-447.0, 19.4, 13.2)$ km s^{-1} and $\boldsymbol{V}_{\mathrm{dn}}=(-317.5, 1.0, -111.6)$ km s^{-1}. With these values, $\boldsymbol{n}_{\mathrm{sh}}=(0.85, -0.02, -0.53)$ in GSE coordinates. Using the four-point, multi-spacecraft timing, the speed of the shock V_{sh} and the shock normal direction $\boldsymbol{n}_{\mathrm{sh}}$ can be determined using

$$\begin{bmatrix}\boldsymbol{r}_{12}\\ \boldsymbol{r}_{13}\\ \boldsymbol{r}_{14}\end{bmatrix}\frac{1}{V_{\mathrm{sh}}}\begin{bmatrix}n_{\mathrm{sh},x}\\ n_{\mathrm{sh},y}\\ n_{\mathrm{sh},z}\end{bmatrix}=\begin{bmatrix}t_{12}\\ t_{13}\\ t_{14}\end{bmatrix}$$

where $\boldsymbol{r}_{xy}$ is the distance vector between spacecraft x and y and t_{xy} is the timing difference of the same boundary observed by both spacecraft. Using this with the four-point observations of MMS crossing the HFA shock, the shock normal direction is calculated as $\boldsymbol{n}_{\mathrm{sh}}=(0.84, -0.44, -0.31)$ in GSE coordinates, which is generally consistent with the result from co-planarity analysis, the speed as $V_{\mathrm{sh}}=-134$ km s^{-1}, which when transformed into the solar-wind frame yields a speed of about 360 km s^{-1} upstream (towards the Sun). As a simple, independent check, these shock normal directions can also be compared to an idealized solution in which the velocity tangential to the shock is conserved, meaning that the difference in velocity from upstream to downstream is approximately along the shock normal. With $\boldsymbol{V}_{\mathrm{up}}$ and $\boldsymbol{V}_{\mathrm{dn}}$ as listed above, this yields a shock normal of $\boldsymbol{n}_{\mathrm{sh}}=(0.72, -0.10, -0.69)$, which again is a rough approximation and is generally in the same sense as the more accurate forms above.

The size of the HFA is calculated using the normalized tracking speed of the HFA along the bow shock[7]. With this and the observed time of transit through the HFA, a size of 2.51 Earth radii is calculated. Finally, the expansion speed of the HFA is calculated using the boundary motion of the upstream and downstream boundaries of the HFA[7], yielding an expansion speed of about 108 km s^{-1}.

Conceptual model for the Fermi acceleration trap. Figure 3 and Extended Data Fig. 2 illustrate the conceptual model of the Fermi acceleration trap that we worked with for this study. In essence, ions within the very low-field-strength core of the HFA can reflect back and forth between the regions of higher field strength at the HFA sheath or shock and Earth's bow shock or magnetosheath. Although the local bow shock may be destroyed in an effective manner adjacent to the HFA, on the basis of simulations there is still a strong magnetic gradient at the transition from the core of the HFA (very low field strength) into the magnetosheath plasma. This magnetic gradient can serve as a mirror point and reflect particles in the acceleration trap. A key aspect here is that in the very low-field-strength field (at most 1 nT) in the core of the HFA, ions with energies of tens of kiloelectronvolts to more than 100 keV have gyroradii that are larger than the length scale of the trap L. Thus, the ion trajectories are essentially ballistic within the core. However, the fields on either side of the HFA core, in the HFA sheath or shock and at Earth's bow shock or magnetosheath are much stronger, roughly 10–30 nT. Thus, on either side of the HFA core, the ions will experience a strong magnetic gradient that is capable of reflecting a subset of the incident particles (depending on their gyrophase and pitch angle, as we account for in the model that is described below). Because the HFA sheath or shock is converging continuously on Earth's bow shock or magnetosheath ($\mathrm{d}L/\mathrm{d}t<0$), this establishes ideal conditions for first-order Fermi acceleration to accelerate some subset of the particles within the trap. Only a small subset of the particles in our model acceleration trap will remain trapped for a long time and attain very high energies. However, because there are exponentially more particles at lower energies, even if a small fraction of these particles are efficiently accelerated up to several hundred kiloelectronvolts in the trap, the result can be quite drastic compared to the usual population at those energies (typically at unobservable or near instrument-background levels). Considering that the ions might experience additional acceleration at the HFA shock or due to reflections from waves within the system adds to the efficiency of the system's acceleration effect.

Maximum energy in the Fermi acceleration trap. From Fermi[14], a formula was previously derived[15] to calculate an accelerated ion's energy at some time $E(t)$, using its initial energy E_0 and the ratio of the initial length scale L_0 of the Fermi acceleration trap (the distance between converging magnetic mirror points) to the length at the same time $L(t)$:

$$E(t)=E_0\left[\frac{L_0}{L(t)}\right]^{2/3} \tag{1}$$

This equation assumes that the speed of the ion is much greater than that of the convergence of the trap ($v\gg U$), which enables higher-order terms to be excluded. Considering that we are studying ions with kinetic energies of about 100 keV to 1 MeV, their speeds of thousands of kilometres per second far exceed the converging speed of about 300 km s^{-1}, and thus the assumption is valid. We also assume that the initial seed population consists of the suprathermal ions in the quasi-parallel foreshock, with energies in the range of tens of kiloelectronvolts to about 150 keV, which also satisfy the required condition. Next, we make two further assumptions: first, that the initial energy of the accelerated ion is proportional to the kinetic energy of the solar wind, $E_0\propto mV_{\mathrm{sw}}^2/2$; and second, that the minimum size of the Fermi acceleration trap that can effectively accelerate an ion is proportional to the ion's gyroradius, $L_{\mathrm{min}}\propto mv_{\mathrm{max}}/(qB)$, where m the mass of the ion, V_{sw} is the solar-wind speed, q is the ion's charge state, B is the magnetic field strength and v_{max} is the limit of the ion's velocity assuming it can escape the Fermi acceleration trap once its gyroradius is some substantial fraction of the trap length scale. Using the assumptions with equation (1), with $E(t)=E_{\mathrm{max}}=mv_{\mathrm{max}}^2/2$ and $L(t)=L_{\mathrm{min}}$, we can easily derive a relationship between the expected maximum energy E_{max}, mass and charge state of an escaping ion for any particular Fermi acceleration trap (that is, L_0, B and V_{sw} are constants for all ion species in the trap):

$$\frac{E_{\mathrm{max}}^2}{mq}=\text{constant} \tag{2}$$

Accelerated ion energy distribution in the Fermi acceleration trap. Because the accelerated ion gyroradii are very large compared to the HFA boundary, during each bounce in the Fermi acceleration trap between the bow shock and HFA shock, a fraction of ions can escape the trap. To obtain the fraction of ions that escape the trap, we develop a new model. The field strength at the HFA boundary is typically much stronger than in the core, so the magnetic field ($\boldsymbol{B}$) at the boundary is approximately along the boundary surface to maintain divergenceless $\boldsymbol{B}$. This is illustrated in Extended Data Fig. 3.

How far into a HFA sheath or shock boundary of thickness d an ion can penetrate is determined by $p=R-R\sin(\varphi)$, with $\varphi\in(-\pi/2, \pi/2]$, where R is the ion's gyroradius and φ is the gyrophase when the ion enters the HFA boundary. When $P>d$, that is, $\sin(\varphi)<1-d/R$, ions escape through the HFA boundary; but for $P<d$ ions are reflected at the boundary and can remain within the Fermi acceleration trap. For a certain R (or certain energy and pitch angle) and assuming a gyrotropic distribution, the fraction of leaked ion flux is

$$X=\frac{1}{2}\int_{-\pi/2}^{\varphi_{\mathrm{c}}}\cos(\varphi)\mathrm{d}\varphi=\frac{\sin(\varphi_{\mathrm{c}})+1}{2}=1-\frac{d}{2R}$$

where φ_c is the critical gyrophase for escape, the cosine function results from the flux of ions that move towards the boundary being proportional to $v_{\mathrm{n}}=v_{\perp}\cos(\varphi)$, and the factor of 1/2 is from normalization.

Next, we can calculate the fraction of leaked ion flux as a function of pitch angle θ. For any ion speed v, ions can leak out only when the pitch angle allows for $R>d$, so there is a critical pitch angle: $\sin(\theta_{\mathrm{c}})=qBd/(mv)$. Assuming an isotropic pitch angle distribution within the trap, the fraction of leaked ion flux as a function of pitch angle is[36]

$$X=\frac{4}{\pi}\int_{\theta_{\mathrm{c}}}^{\pi/2}\left[1-\frac{\sin(\theta_{\mathrm{c}})}{2\sin(\theta)}\right]\sin^2(\theta)\mathrm{d}\theta=1-\frac{2\theta_{\mathrm{c}}}{\pi}$$

With this function, X linearly decreases from 1 to 0 as θ_c increases from 0 to $\pi/2$. Note that θ_c is a function of charge, mass and momentum.

This model can now be used to calculate the energy spectra of ions accelerated within the Fermi trap. Here, we make the assumption that the ions are always reflected by the bow shock without any substantial energy gain; that is, we focus on how multiple reflections at the converging HFA shock affect the ion energy spectrum when considering this model for accelerated ion escape.

After some number n of reflections within the Fermi acceleration trap, the speed of an accelerated ion is

$$v_n^2 = v_{n-1}^2 + 4v_{n-1}U\cos(\alpha) + 4U^2$$

where U is the convergence speed of the HFA shock upon the bow shock and α is the angle between the injection speed and the boundary normal. Assuming an isotropic particle distribution, on average

$$v_n^2 = v_{n-1}^2 + \frac{8}{3}v_{n-1}U + 4U^2$$

Next, the probability of an ion reflecting n times within the trap is

$$P_n = \prod_{i=0}^{n-1}(1-X_i) = \prod_{0}^{n-1}\frac{2\theta_{ci}}{\pi}$$

The normalized ion distribution function f can be approximated as $f \approx \partial P/\partial v$, and the result using this model must be calculated numerically. Guided by the MMS observations, we used $d = 1{,}000$ km and $U = 220$ km s^{-1} over $n = 20$ bounces to obtain the distributions shown in Fig. 4.

We compared these results with those when also including escape on the downstream side (away from the Sun) of the bow shock. To do this, we used a model[37] to include loss of particles into the magnetosheath. In this model, the probability of escape to the sheath is $P = 4U_{\rm sh}/v$, where $U_{\rm sh}$ is the speed of plasma in the magnetosheath (around 100 km s^{-1}) and v is the particle velocity, such that most particles cross the bow shock many times before escaping downstream. For ions with energies of more than about 100 keV, the probability of escape from the trap is much, much higher in the model developed here than for the bow-shock model. The bow-shock-escape model results in a power-law distribution of the escaping particles, and when included with the model developed here contributes a negligible amount of loss that still results in an exponential distribution for the accelerated ions escaping from the trap, like those shown in Fig. 4 for the three different ion species.

Estimating the length of the Fermi acceleration trap. Using the observed energy gain of the accelerated ions, we can estimate the length scale between the HFA shock and bow shock assuming that the ions were accelerated within a first-order Fermi trap between those two boundaries. With the average energy gain from a particle bouncing between the two converging shocks and assuming isotropic scattering, the average energy gain of the particle is $\langle \Delta E \rangle = 4mvU/3$. Using this, along with the particle's bounce frequency in the trap, $f_b = \langle v \rangle/(2L) = v/(4L)$, where $\langle v \rangle$ is the average velocity accounting for pitch angle and L is the length scale of the trap, we can approximate the average rate of energy gain:

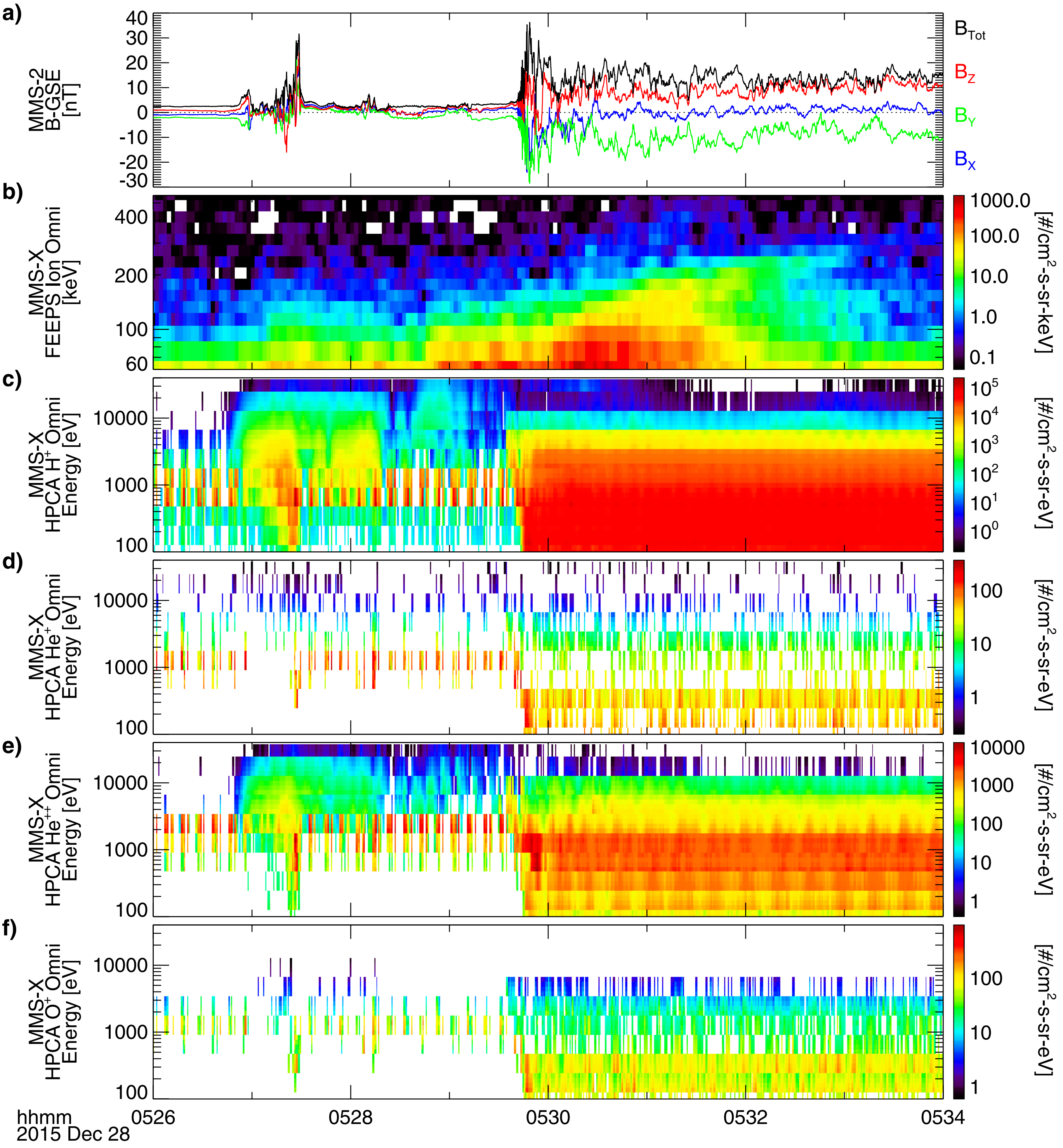
$$\frac{dE}{dt} = \frac{mv^2U}{3L} \tag{3}$$

Using α particles (which allow multiple data points over a broad energy range, unlike protons, and an exact mass, unlike the CNO ions) with an energy of about 200 keV, we can estimate the length scale using the following observed quantities: $dE/dt \approx 220$ keV/(180 s) $= 1.96 \times 10^{-16}$ J s^{-1}, $v \approx 3 \times 10^6$ m s^{-1} (for a 200-keV α particle), $U \approx 200$ km s^{-1} (approximation of HFA shock expansion speed plus motion of the bow shock towards the Sun from multi-spacecraft timing during the crossing) and $m = 6.6951 \times 10^{-27}$ kg. Using these values in equation (3) yields a length scale of about 20,516 km = 3.2 Earth radii, which is consistent with the HFA scale observed by MMS (about 2.5 Earth radii). Note that this is only an order-of-magnitude scale comparison: the size of the HFA observed by MMS depends on where and how the HFA passed over the spacecraft, whereas the length scale calculated from theory is the most effective or representative length within the acceleration trap between the HFA and the bow shock. We expect the two to be of the same order for scale (that is several Earth radii), which is exactly as we have calculated here.

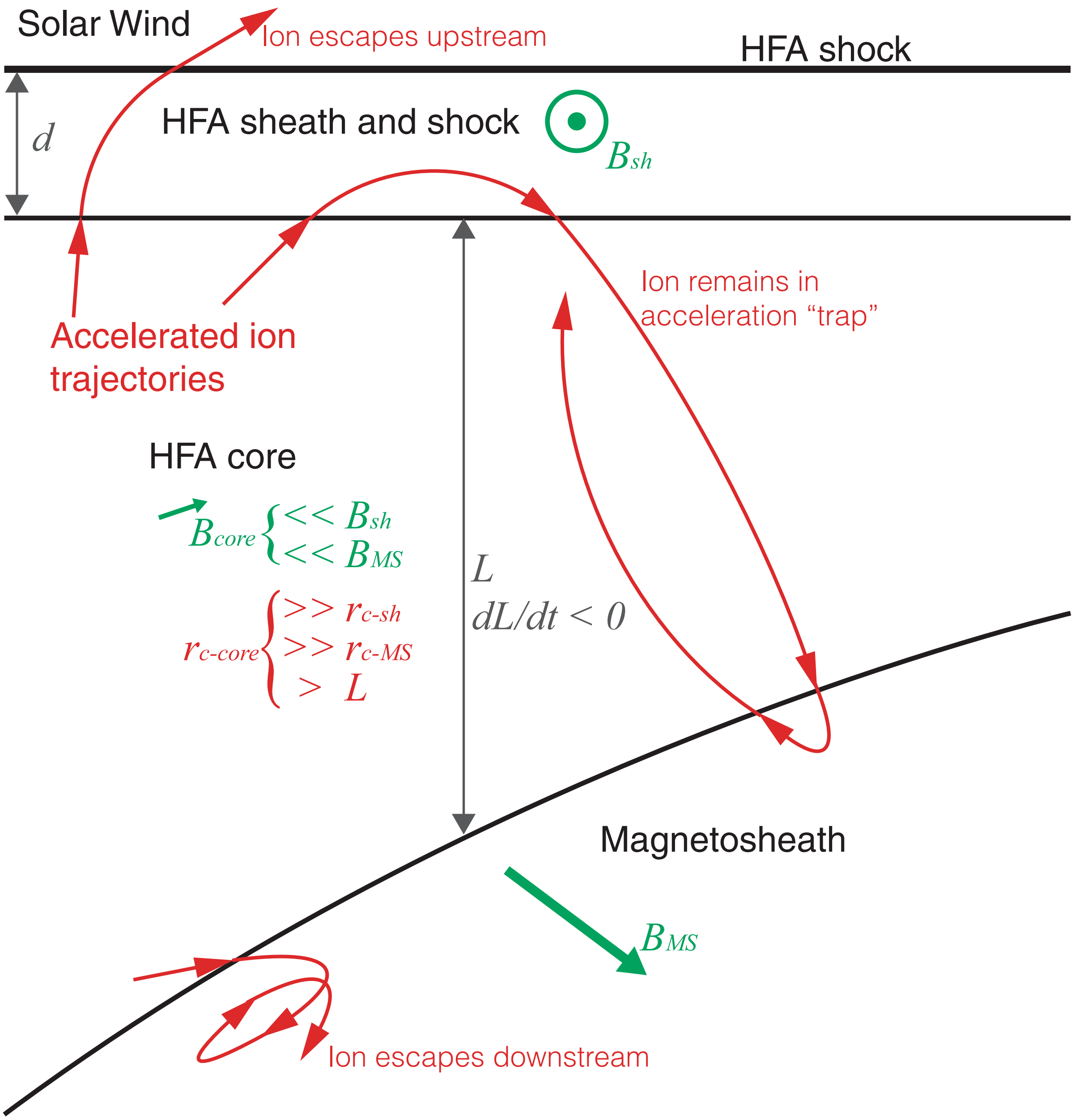
Code availability. All code used to analyse the MMS data in this study is based on the publicly available SPEDAS tools (http://themis.ssl.berkeley.edu/software.shtml). The code and theoretical results used for model–data comparisons will be made available on request to the corresponding author.

Data availability. All MMS data used for this study are publicly available via the MMS Science Data Center at https://spdf.gsfc.nasa.gov/pub/data/mms.

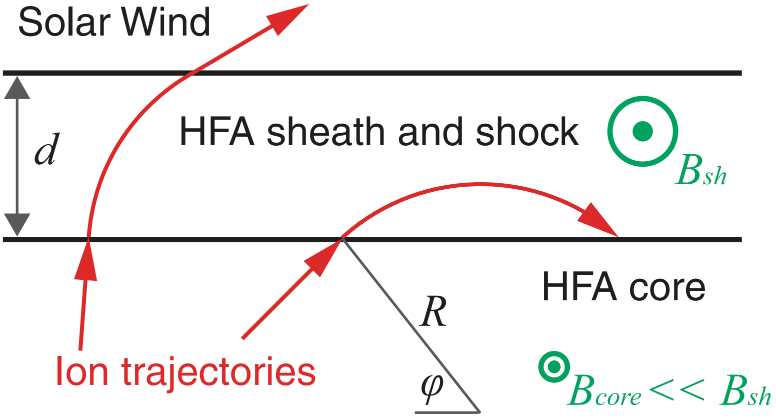
29. Torbert, R. B. et al. The FIELDS instrument suite on MMS: scientific objectives, measurements, and data products. *Space Sci. Rev.* **199**, 105–135 (2016).
30. Mauk, B. H. et al. The Energetic Particle Detector (EPD) investigation and the Energetic Ion Spectrometer (EIS) for the Magnetospheric Multiscale (MMS) mission. *Space Sci. Rev.* **199**, 471–514 (2016).
31. Blake, J. B. et al. The Fly's Eye Energetic Particle Spectrometer (FEEPS) sensors for the Magnetospheric Multiscale (MMS) mission. *Space Sci. Rev.* **199**, 309–329 (2016).
32. Pollock, C. et al. Fast plasma investigation for Magnetospheric Multiscale. *Space Sci. Rev.* **199**, 331–406 (2016).
33. Young, D. T. et al. Hot plasma composition analyzer for the Magnetospheric Multiscale mission. *Space Sci. Rev.* **199**, 407–470 (2016).
34. Sonnerup, B. U. Ö. & Scheible, M. in *Analysis Methods for Multi-Spacecraft Data* (eds Paschmann, G. & Daly, P. W.) Ch. 8, 185–220 (ISSI/ESA, 1998).
35. Schwartz, S. J. in *Analysis Methods for Multi-Spacecraft Data* (eds Paschmann, G. & Daly, P. W.) Ch. 10, 249–270 (ISSI/ESA, 1998).
36. Liu, T. Z., Lu, S., Angelopoulos, V., Lin, Y. & Wang, X. Y. Ion acceleration inside foreshock transients. *J. Geophys. Res.* **123**, 163–178 (2018).
37. Drury, L. O. An introduction to the theory of diffusive shock acceleration of energetic particles in tenuous plasmas. *Rep. Prog. Phys.* **46**, 973–1027 (1983).



Extended Data Fig. 1 | Solar-wind ion species observed by MMS HPCA. **a**, Magnetic-field vector and strength from MMS-2. **b**–**f**, FEEPS total energetic ion (**b**) and HPCA ion (**c**–**f**) composition flux energy spectra (omni-directional) combined from all four spacecraft (MMS-X). HPCA observed many fewer He^+ ion than He^{2+} ions (α particles) and no O^+ ions (all zero counts despite energy channels up to about 38 keV) at $E > 10$ keV during the observation of the energetic ions between the HFA and the magnetosheath. The absence of low-charge-state oxygen in HPCA data (O^+, O^{2+}) indicates that the energetic CNO ions observed by EIS were at a high charge state. This confirms that the accelerated ions were high-charge-state ions ($q/m \geq 2$) and thus of solar-wind origin. Like FEEPS, HPCA also shows that the most energetic beam of α particles and protons is streaming away from the HFA.


Solar Wind
Ion escapes upstream
HFA shock
d
HFA sheath and shock
B_{sh}
Ion remains in acceleration “trap”
Accelerated ion trajectories
HFA core
$B_{core} \begin{cases} << B_{sh} \\ << B_{MS} \end{cases}$
$r_{c\text{-}core} \begin{cases} >> r_{c\text{-}sh} \\ >> r_{c\text{-}MS} \\ > L \end{cases}$
L
$dL/dt < 0$
Magnetosheath
B_{MS}
Ion escapes downstream

Extended Data Fig. 2 | Conceptual model of the Fermi acceleration trap created by an HFA evolving along Earth’s bow shock. At the top of this schematic is the solar wind, which is the ‘upstream’ direction to which some ions within the trap might escape. The region at the bottom of the schematic is the magnetosheath plasma, which is the ‘downstream’ direction to which some ions might also escape. The HFA sheath and shock region, with some finite thickness *d*, lies between the two horizontal black lines. The HFA core region lies between the HFA sheath/shock and magnetosheath regions. The HFA core has a characteristic length scale *L* and is perpetually converging on the magnetosheath such that $dL/dt < 0$ at any point along the core. Critical to the model (and as observed and simulated in HFAs), the magnetic-field strength within the HFA core (less than 1 nT) is much, much weaker than that in the HFA sheath or shock and in Earth’s bow shock or magnetosheath (more than 10 nT); this geometry ensures that some subset of ions within the HFA core can remain trapped between the two regions of higher field strength, owing to differences in the ion gyroradii (r_c) in the different regions. We note that the orientations of the magnetic fields are sketched arbitrarily here. Three example ion trajectories are sketched: (i) escaping to the solar wind; (ii) escaping to the magnetosheath; and (iii) remaining trapped within the HFA core.



Extended Data Fig. 3 | Simplified model of the HFA. This model includes a finite thickness of the HFA sheath (that is, shocked plasma) and shock *d*, which is used to determine the energy spectra of ions accelerated within the trap considering losses through the boundary. In this picture, the HFA sheath and shock fall between the two horizontal black lines, with the HFA core being a larger region below this and the solar wind being the region above this. Only ions at particular energies, pitch angles and gyrophase can remain trapped in the Fermi acceleration trap between Earth's bow shock or magnetosheath and the HFA shock. For example, the two ion trajectories shown are at the same energy but different gyrophase when they enter the HFA sheath with enhanced magnetic field strength: the rightmost ion is reflected and can continue to be accelerated, whereas the leftmost ion escapes the trap and is lost to the solar wind.

LETTER

OPEN

https://doi.org/10.1038/s41586-018-0435-1

Observation of the 1S–2P Lyman-α transition in antihydrogen

M. Ahmadi[1], B. X. R. Alves[2], C. J. Baker[3], W. Bertsche[4,5], A. Capra[6], C. Carruth[7], C. L. Cesar[8], M. Charlton[3], S. Cohen[9], R. Collister[6], S. Eriksson[3], A. Evans[10], N. Evetts[11], J. Fajans[7], T. Friesen[2,10], M. C. Fujiwara[6]*, D. R. Gill[6], J. S. Hangst[2]*, W. N. Hardy[11], M. E. Hayden[12], E. D. Hunter[7], C. A. Isaac[3], M. A. Johnson[4,5], J. M. Jones[3], S. A. Jones[2,3], S. Jonsell[13], A. Khramov[6], P. Knapp[3], L. Kurchaninov[6], N. Madsen[3], D. Maxwell[3], J. T. K. McKenna[6], S. Menary[14], J. M. Michan[6,15], T. Momose[11,16]*, J. J. Munich[12], K. Olchanski[6], A. Olin[6,17], P. Pusa[1], C. Ø. Rasmussen[2], F. Robicheaux[18], R. L. Sacramento[8], M. Sameed[4], E. Sarid[19], D. M. Silveira[8], D. M. Starko[14], G. Stutter[2], C. So[10], T. D. Tharp[20], R. I. Thompson[6,10], D. P. van der Werf[3,21] & J. S. Wurtele[7]

In 1906, Theodore Lyman discovered his eponymous series of transitions in the extreme-ultraviolet region of the atomic hydrogen spectrum[1,2]. The patterns in the hydrogen spectrum helped to establish the emerging theory of quantum mechanics, which we now know governs the world at the atomic scale. Since then, studies involving the Lyman-α line—the 1S–2P transition at a wavelength of 121.6 nanometres—have played an important part in physics and astronomy, as one of the most fundamental atomic transitions in the Universe. For example, this transition has long been used by astronomers studying the intergalactic medium and testing cosmological models via the so-called 'Lyman-α forest'[3] of absorption lines at different redshifts. Here we report the observation of the Lyman-α transition in the antihydrogen atom, the antimatter counterpart of hydrogen. Using narrow-line-width, nanosecond-pulsed laser radiation, the 1S–2P transition was excited in magnetically trapped antihydrogen. The transition frequency at a field of 1.033 tesla was determined to be 2,466,051.7 ± 0.12 gigahertz (1σ uncertainty) and agrees with the prediction for hydrogen to a precision of 5×10^{-8}. Comparisons of the properties of antihydrogen with those of its well-studied matter equivalent allow precision tests of fundamental symmetries between matter and antimatter. Alongside the ground-state hyperfine[4,5] and 1S–2S transitions[6,7] recently observed in antihydrogen, the Lyman-α transition will permit laser cooling of antihydrogen[8,9], thus providing a cold and dense sample of anti-atoms for precision spectroscopy and gravity measurements[10]. In addition to the observation of this fundamental transition, this work represents both a decisive technological step towards laser cooling of antihydrogen, and the extension of antimatter spectroscopy to quantum states possessing orbital angular momentum.

Challenges to antimatter Lyman-α spectroscopy include the difficulty of fabricating optical components and continuous-wave[11] or pulsed laser sources at these extremely short wavelengths, as well as the scarcity of anti-atoms. The current observation was made possible

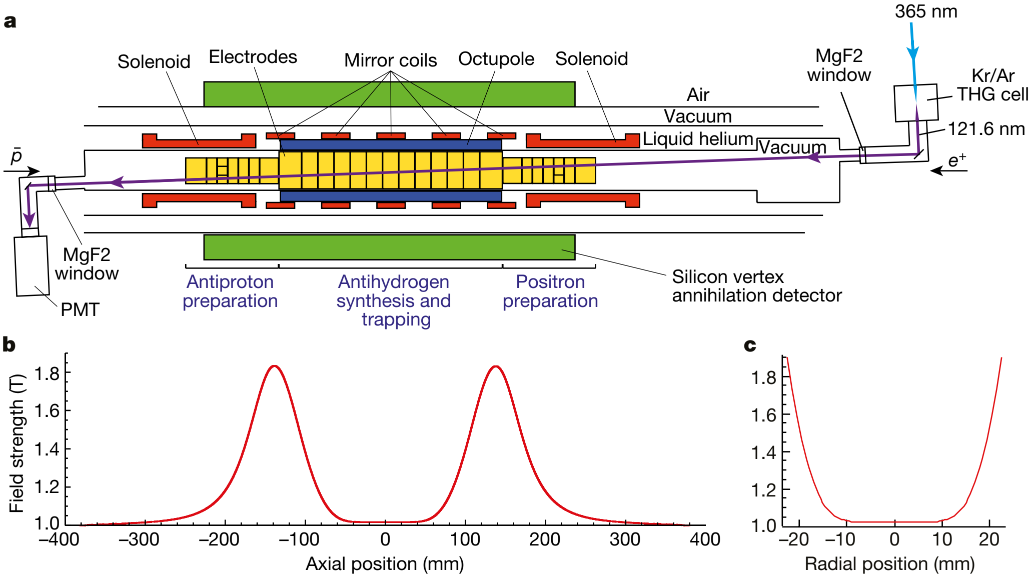


Fig. 1 | Experimental set-up. **a**, The three-layer silicon vertex annihilation detector is shown schematically in green; the external solenoid magnet for the Penning traps is not shown in this diagram. Laser light enters from the positron (e^+) side (right) and is transmitted to the antiproton ($\bar{p}$) side (left) through vacuum-ultraviolet-grade MgF_2 ultrahigh-vacuum windows. The laser beam crosses the trap axis at an angle of 2.3°. The transmitted 121.6-nm pulses are detected by a photomultiplier at the antiproton side. **b**, Axial magnetic well formed by the five mirror coils and responsible for the axial confinement of cold (less than 0.5 K) anti-atoms. **c**, Radial magnetic octupole field profile. PMT, photomultiplier tube; THG, third-harmonic generation.

---

[1]Department of Physics, University of Liverpool, Liverpool, UK. [2]Department of Physics and Astronomy, Aarhus University, Aarhus, Denmark. [3]Department of Physics, College of Science, Swansea University, Swansea, UK. [4]School of Physics and Astronomy, University of Manchester, Manchester, UK. [5]Cockcroft Institute, Sci-Tech Daresbury, Warrington, UK. [6]TRIUMF, Vancouver, British Columbia, Canada. [7]Department of Physics, University of California at Berkeley, Berkeley, CA, USA. [8]Instituto de Fisica, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil. [9]Department of Physics, Ben-Gurion University of the Negev, Beer-Sheva, Israel. [10]Department of Physics and Astronomy, University of Calgary, Calgary, Alberta, Canada. [11]Department of Physics and Astronomy, University of British Columbia, Vancouver, British Columbia, Canada. [12]Department of Physics, Simon Fraser University, Burnaby, British Columbia, Canada. [13]Department of Physics, Stockholm University, Stockholm, Sweden. [14]Department of Physics and Astronomy, York University, Toronto, Ontario, Canada. [15]École Polytechnique Fédérale de Lausanne (EPFL), Swiss Plasma Center (SPC), Lausanne, Switzerland. [16]Department of Chemistry, University of British Columbia, Vancouver, British Columbia, Canada. [17]Department of Physics and Astronomy, University of Victoria, Victoria, British Columbia, Canada. [18]Department of Physics and Astronomy, Purdue University, West Lafayette, IN, USA. [19]Soreq NRC, Yavne, Israel. [20]Physics Department, Marquette University, Milwaukee, WI, USA. [21]IRFU, CEA/Saclay, Gif-sur-Yvette Cedex, France. *e-mail: Makoto.Fujiwara@triumf.ca; jeffrey.hangst@cern.ch; momose@chem.ubc.ca

by a number of technical advances, including the development of a solid-state-based, pulsed Lyman-α source[12], implementation of innovative plasma control techniques[13], and the ALPHA Collaboration's recent, marked improvement in antihydrogen trapping and accumulation or 'stacking' rate. Stacking provides a sample of several hundred anti-atoms[14], accumulated over several hours. Taking advantage of a nanosecond-scale laser pulse and our low-background annihilation detection, we have also inferred information on the antihydrogen velocity distribution.

Because matter and antimatter annihilate each other when they meet, antihydrogen must be created and then trapped in strong, inhomogeneous magnetic fields in an ultrahigh-vacuum chamber. The ALPHA-2 apparatus (Fig. 1a) is designed to combine antiprotons from CERN's Antiproton Decelerator[15] with positrons from a positron accumulator[16,17] to produce and to trap atoms of antihydrogen. The techniques that we have developed to produce antihydrogen cold enough (below 0.54 K) to trap can be found elsewhere[13,14,18,19]. A typical trapping trial in ALPHA-2 involves mixing about 90,000 antiprotons and three million positrons to produce 50,000 antihydrogen atoms. Most of the antihydrogen atoms produced are either too energetic or in the wrong quantum state to be captured; typically 10–20 anti-atoms are trapped in one four-minute-long mixing cycle. We have employed the stacking technique referred to above to increase the total number of trapped antihydrogen atoms to several hundred for each measurement sequence.

The trapped anti-atoms are confined by the interaction of their magnetic moments with the inhomogeneous magnetic field, shown in Fig. 1b, c. The cylindrical trapping volume for antihydrogen has a diameter of 44.35 mm and an axial length of 280 mm. The magnetic field minimum of 1.033 T is near the centre of the trap, where the field is oriented axially. The field strength is flat to within one part in 10,000 in a 60 mm (axial) by 4.5 mm (radial) volume around the centre of the trap. The maximum field differential in the trap is 0.82 T along the axial direction and about 0.85 T along the radial direction.

Figure 2 shows the energy levels of hydrogen in the 1S and 2P states in a magnetic field. The 2S state is also depicted for reference. Antihydrogen atoms in the states labelled as 1S$_\mathrm{c}$ and 1S$_\mathrm{d}$ can be trapped. At zero magnetic field, the excited 2P state splits into two states (2P$_{3/2}$ and 2P$_{1/2}$) owing to the relativistic spin–orbit interaction. In non-zero magnetic fields, due to the Zeeman effect, the 2P$_{3/2}$ state splits into four sublevels, while the 2P$_{1/2}$ state splits into two. Each of these sublevels in turn splits into two states owing to the nuclear spin of the antiproton, but the splitting is much smaller (less than 0.1 GHz) than that in the ground (1S) state, and is therefore not visible on the scale of Fig. 2.

The 1S–2P$_\mathrm{c}$ transition studied here is a dipole-allowed transition. When antihydrogen is excited to one of the 2P sublevels, the excited state decays to the ground 1S state within a few nanoseconds by emitting a photon at 121.6 nm. Owing to the mixing of the positron spin states, there is a non-zero probability that the excited atom decays to either the 1S$_\mathrm{a}$ or 1S$_\mathrm{b}$ state instead of the original 1S$_\mathrm{c}$ or 1S$_\mathrm{d}$ state. Those that decay to the 1S$_\mathrm{a}$ or 1S$_\mathrm{b}$ state escape and annihilate at the trap walls and can be detected by the ALPHA-2 silicon vertex detector. The silicon vertex detector affords us single-atom detection capability, which is key to antimatter spectroscopy with few atoms[4,5]. The silicon vertex detector tracks the charged pions from the antiproton annihilation; the pion tracks are reconstructed to determine the location (vertex) of each annihilation. We employ machine-learning techniques, incorporating several parameters from the detector and the track reconstruction in a multi-variate analysis[20] (MVA; see Methods), to distinguish antiproton annihilations from the cosmic ray background.

In this experiment, the antihydrogen atoms were excited to the 2P$_\mathrm{c}$ state (Fig. 2) by light that was linearly polarized, the polarization vector being nearly perpendicular to the direction of the axial magnetic field. Narrow-line-width (about 65 MHz), pulsed (about 12-ns duration) laser radiation at 121.6 nm was used for the excitation. The system is described in detail in Methods, and the schematic diagram of the laser system is shown in Extended Data Fig. 1.

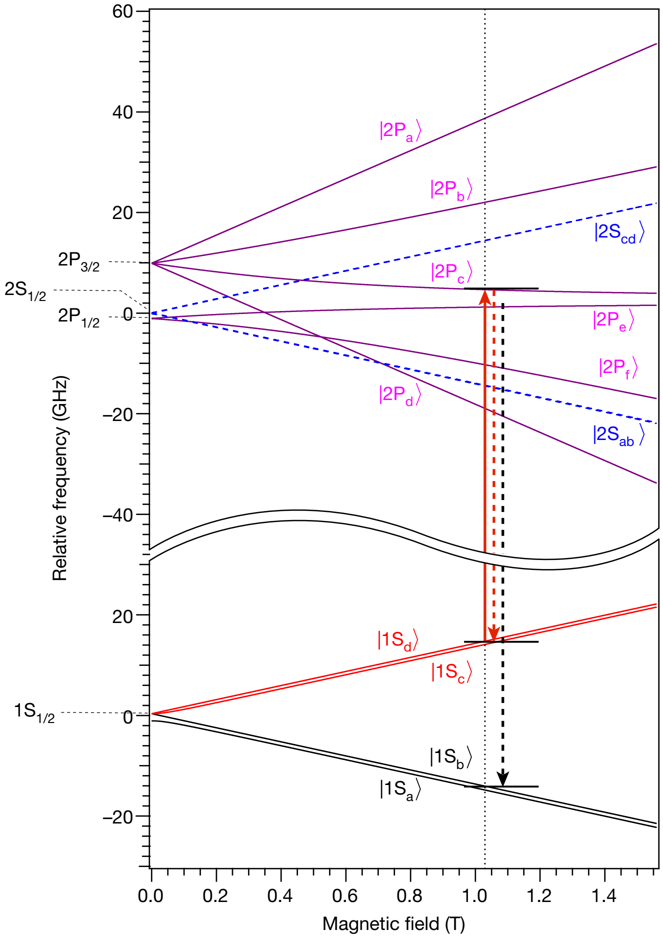


Fig. 2 | Trappable and untrappable energy levels in hydrogen. This plot reports the calculated energies for the sublevels of the 1S, 2S and 2P states for hydrogen as functions of the magnetic field strength. Note that the centroid energy difference $E_{1\mathrm{S}-2\mathrm{S}} = 2.4661 \times 10^{15}$ Hz is suppressed on the vertical axis. The vertical red arrow indicates the one-photon laser transition addressed here; the dashed red arrow illustrates the decay to the same trappable level, whereas the dashed black arrow corresponds to the decay to the untrappable level. Antihydrogen that decays to the untrapped 1S level escapes from the trap and can be detected by the annihilation detector.

For each experimental sequence, about 500 antihydrogen atoms were accumulated in the trapping region by multiple stacking (typically 30 stacks) over an approximately two-hour period. Experience from past microwave manipulation experiments[5,6] indicates that the 1S$_\mathrm{c}$ and 1S$_\mathrm{d}$ states are equally populated in the initially trapped sample. The trapped atoms were irradiated for about two hours by laser pulses using a 10-Hz repetition rate. During each two-hour sequence, the trapped atoms were exposed to radiation at twelve different frequencies around the calculated hydrogen resonance frequency of 2,466,051.625 GHz (using the average of the two 1S hyperfine levels). The magnetic field magnitude was determined by electron cyclotron resonance[21]. The frequency detunings ranged from −2.47 GHz to +0.85 GHz. The laser frequency was switched every 20 s in a pattern designed to minimize saturation and depletion effects (due to the finite population size) at each frequency. After each two-hour exposure, the remaining antihydrogen atoms were released and counted by ramping down the trap magnets in 15.6 s. We repeated the described sequence four times. The number of detected events is summarized in Table 1. With average laser pulse energies between 0.53 nJ and 0.65 nJ, about 60% of the trapped antihydrogen atoms were expelled and detected during the exposure time of two hours.

Because the laser is pulsed, and the excitation light is present for about 12 ns for each pulse, the annihilation events due to the excitation

Table 1 | Summary of experimental data

Sequence	Average pulse energy (pJ)	Number of pulses at each frequency	Total detected events during laser irradiation	Total detected events from the release of remaining atoms
1	620	6,000	230	170
2	620	6,000	254	164
3	650	6,000	261	179
4	530	6,000	221	174

The total number of antihydrogen atoms detected during the laser irradiation and during the release of remaining atoms is tabulated for each sequence. The MVA identifies annihilations with an efficiency of 0.807 for the laser irradiation and 0.851 for the release of the remaining atoms.

are only expected to occur in the approximately 1-ms time window in which the untrapped atoms are forced to the trap wall. This greatly helps us to distinguish between signals due to laser-driven annihilations and those due to cosmic background. Figure 3 illustrates the power of the pulsed excitation method by comparing (1) on-resonance and (2) off-resonance measurements of the so-called time-of-flight (TOF) distribution. Even without the sophisticated machine-learning algorithm, raw detector triggers (Methods, Fig. 3a, black line) show a high signal level for the on-resonant case, in the time window between 0 ms and 1 ms after the laser pulse. The background is approximately constant. With the MVA (red line) the background almost disappears. In Fig. 3c, d we present two-dimensional plots for the axial position versus the TOF of the annihilation events. These plots also illustrate the selection criteria or 'cuts' (0 ms $< t <$ 1 ms and -100 mm $< z <$ 100 mm) used to accept events for the spectral shape determination described below. These cuts were predetermined on the basis of simulation information.

Figure 4a shows the spectral line shape of the transition, obtained using events with the above cuts applied. In this plot, the probability is normalized to the pulse energy; we have thus assumed that the excitation probability is linearly proportional to the laser pulse energy. The plotted data show that the transition from the $1S_c$ and $1S_d$ states to the $2P_c$ state has a line width of about 1.5 GHz (full-width at half-maximum, FWHM). If the hyperfine coupling constants of antihydrogen between the 1S and 2P states are the same as those for normal hydrogen, the spectrum in the magnetic field of 1.033 T would comprise two lines with a separation of 0.74 GHz. The two distinct lines are caused by the possible orientations of the antiproton nuclear spin in the 1S ground state. However, this hyperfine splitting is not resolved under the present experimental conditions owing to the Doppler broadening of the transitions caused by the motion of trapped antihydrogen atoms parallel with the laser beam. We note that the Doppler shift of the Lyman-α transition of hydrogen with a velocity of 75 m s^{-1} (characteristic for our magnetic trap depth) is 0.6 GHz. A fit to the measured data yields a resonant frequency of 2,466,051.7(0.12) GHz, in good agreement with the calculated hydrogen frequency quoted above. The contributions to the quoted uncertainty of 0.12 GHz are the wavemeter accuracy (0.06 GHz), the 730-nm laser (Methods) cavity lock stability (0.06 GHz), modelling uncertainties (0.07 GHz), and the statistical uncertainty in the curve fit (0.04 GHz). Many of these uncertainties may be reduced in the future, but the large natural line width of $2\pi \times 99.6$ MHz is an obstacle to using the Lyman-α transition for very high-precision measurements as tests of charge–parity–time-reversal (CPT) invariance.

The kinetic energy distribution of the trapped anti-atoms is also of interest. We expect, from previous measurements and simulations of the antihydrogen production and trapping processes[22], that the trapped antihydrogen atoms are not in thermal equilibrium. Our standard model for antihydrogen formation is that the atoms are formed from antiprotons in equilibrium with the positron cloud, which has a temperature of about 20 K. The expected distribution of trapped atoms thus comprises the low-temperature tail of a Maxwellian distribution, truncated at the maximum trap depth of about 0.5 K. The measured line shape of the transition should thus be different from a Gaussian shape for a Doppler-broadened line. Therefore, we carried out detailed simulations (Methods), based on the known physics of hydrogen, to determine the expected line shape and intensity of the transition in our experimental environment. The simulated line shape is model dependent. Two examples of simulated results are shown in Fig. 4a in red and in blue. The calculated transition probability and line width match the data reasonably well over the region measured. The general agreement between the simulation and the observed spectral feature indicates that the observed line shape is indeed dominated by Doppler broadening, corresponding to an average energy of a few hundred millikelvin.

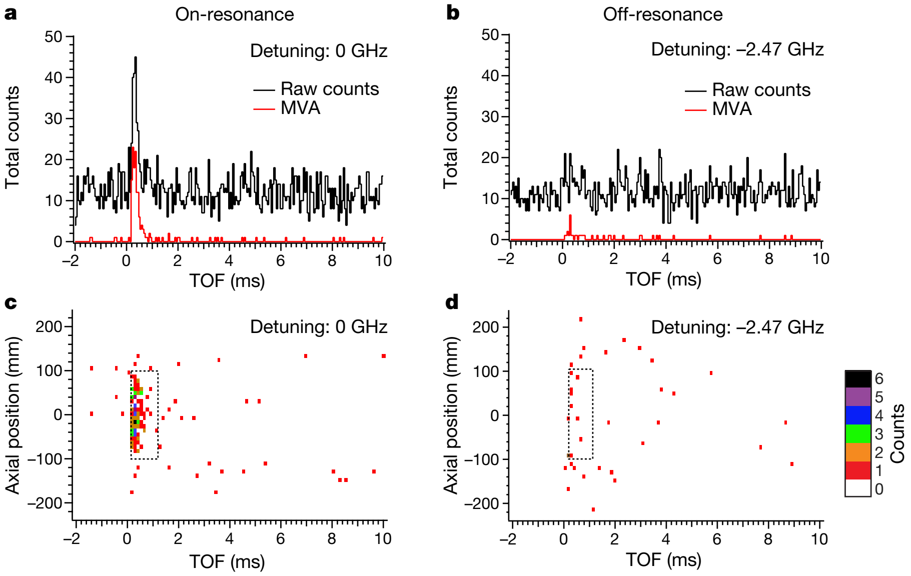


Fig. 3 | Temporal and spatial distribution of detected events. **a**, Temporal distribution of detected signals between -2 ms and 10 ms with respect to the laser pulse. The laser frequency was at the resonance of the 1S–2P transition (detuning 0 GHz). The black trace shows the raw detector triggers. The red trace shows events after background suppression due to an MVA-based machine-learning algorithm. The y axis is the total number of counts (bin size, 50 μs) for 24,000 pulses at the resonant frequency. **b**, Same as **a**, but the laser frequency was off-resonance; the detuning was -2.47 GHz. **c**, The axial position of the annihilation events is plotted versus the TOF. The detuning was 0 GHz. **d**, Same as **c**, but the detuning was -2.47 GHz. Events within the dashed box for each laser frequency were selected for the spectral shape determination (Fig. 4a).

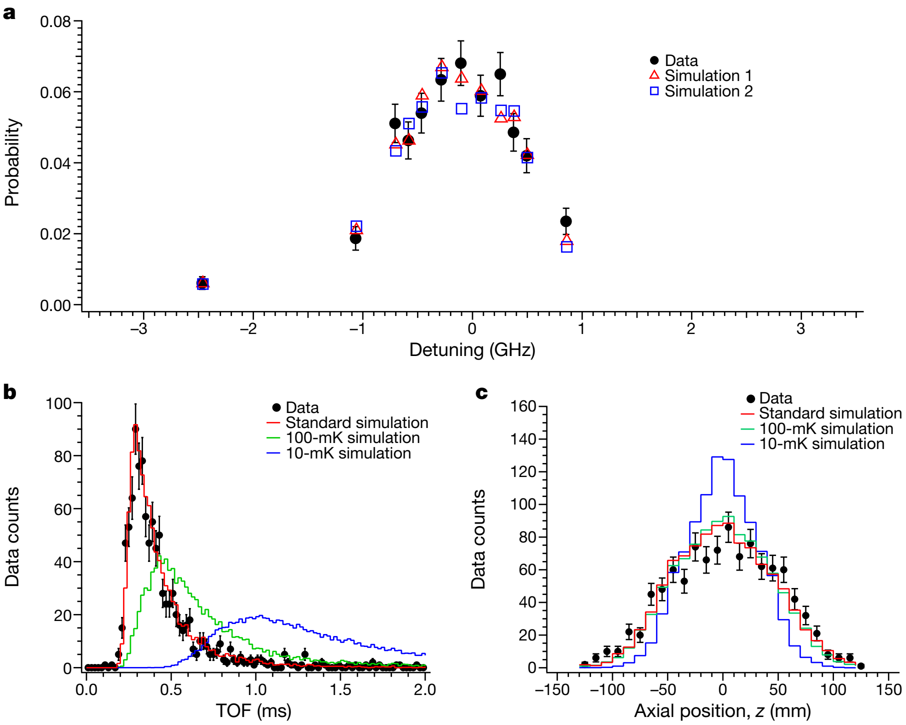


Fig. 4 | 1S–2P line shape, time and axial distributions. a, The 1S–2P spectral line shape. Black, detected events, normalized to the total number of trapped antihydrogen atoms and scaled to the result obtained for a 600-s irradiation at 0.55 nJ. The error bars represent the statistical counting uncertainties. Red and blue, simulated line shapes for the initial conditions of $n = 1$ (simulation 1, red) and $n = 30$ (simulation 2, blue) (Methods). The statistical uncertainties of the simulation are within the size of each marker symbol. **b**, Black, the TOF distribution of all the detected annihilation events, measured relative to the laser pulse. Red, the simulated TOF distribution reflecting the experimental condition. Green and blue, simulated distributions for trapped antihydrogen atoms in thermal equilibrium at 100 mK and 10 mK, respectively. The simulated curves are normalized by area to the data (black). **c**, Comparison of the measured and simulated axial position distributions of the annihilations. The simulated distributions include the detector response function. Black, measured distribution. Red, simulated distribution. Green and blue, simulations with thermal distributions of 100 mK and 10 mK, respectively.

Figure 4b compares the time distribution of the detected signals following the laser pulse (black) to a simulated distribution (red) obtained using the standard formation model. The agreement between the simulation and the detected signal for these TOF distributions confirms that the radial velocity distribution of trapped antihydrogen is consistent with the model. For comparison, we have also simulated the TOF distribution of ejected atoms originating from hypothetical thermal distributions of 100 mK and 10 mK (Fig. 4b), which indicates that the detected time distribution should be sensitive to the radial temperature of trapped antihydrogen. In Fig. 4c we plot the measured axial distribution of laser-induced annihilation events, and we compare it to simulated results at various temperatures. These distributions also exhibit some dependence on antihydrogen temperature and may facilitate diagnosis of future laser cooling attempts.

In conclusion, we have reported here the observation of the 1S–2P, Lyman-α transition in antihydrogen, based on 966 detected events and an estimated background of 14 events. The frequency of this fundamental anti-atomic transition is determined to a precision of about 5×10^{-8}, via narrow-line-width, nanosecond-pulsed, vacuum-ultraviolet laser spectroscopy. We also report a method of directly characterizing the kinetic energy of anti-atoms from their TOF to annihilation, following the laser-induced transition. These observations represent very important steps in the field of low-energy antimatter studies[23–27]. The techniques of optical manipulations and laser cooling, which have revolutionized the field of atomic physics over the past few decades, are about to be applied to anti-atoms. With its natural line width of about 100 MHz (comparable to our laser width), the Lyman-α transition can in principle be employed to cool antihydrogen to the 2.4-mK Doppler limit. Our simulations predict that cooling to about 20 mK is possible with the current ALPHA-2 set-up[9]. This, combined with other planned improvements, would reduce the 1S–2S transition line width by more than an order of magnitude and should eventually allow various other spectroscopic measurements with precisions approaching those achieved in hydrogen[28–30]. At such levels of precision, antihydrogen spectroscopy will have an impact on the determination of fundamental constants[31], in addition to providing elegant tests of CPT symmetry. Furthermore, laser cooling will be crucial for a precision test of the weak equivalence principle via antihydrogen free fall[10] or anti-atom interferometry[32] at the 10^{-2} level and beyond. Access to the 2P state in antihydrogen greatly expands the future experimental horizon; for example, we will now be able to study fine-structure effects in an anti-atom.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0435-1.

Received: 13 March 2018; Accepted: 5 July 2018;
Published online 22 August 2018.

1. Lyman, T. The spectrum of hydrogen in the region of extremely short wave-length. *Mem. Am. Acad. Arts Sci. New Ser.* **13**, 125–146 (1906).
2. Lyman, T. An extension of the spectrum in the extreme ultra-violet. *Nature* **93**, 241 (1914).
3. Weinberg, D. H., Davé, R., Katz, N. & Kollmeier, J. A. The Lyman-α forest as a cosmological tool. *AIP Conf. Proc.* **666**, 157–169 (2003).
4. Amole, C. et al. Resonant quantum transition in trapped antihydrogen atoms. *Nature* **483**, 439–443 (2012).
5. Ahmadi, M. et al. Observation of the hyperfine spectrum of antihydrogen. *Nature* **548**, 66–69 (2017).
6. Ahmadi, M. et al. Observation of the 1S–2S transition in trapped antihydrogen. *Nature* **541**, 506–510 (2017).
7. Ahmadi, M. et al. Characterisation of the 1S–2S transition in antihydrogen. *Nature* **557**, 71–75 (2018).

8. Setija, I. D. et al. Optical cooling of atomic hydrogen in a magnetic trap. *Phys. Rev. Lett.* **70**, 2257–2260 (1993).
9. Donnan, P. H., Fujiwara, M. C. & Robicheaux, F. A proposal for laser cooling antihydrogen atoms. *J. Phys. B* **46**, 025302 (2013).
10. Bertsche, W. A. Prospects for comparisons of matter and antimatter gravitation with ALPHA-g. *Phil. Trans. R. Soc. Lond. A* **376**, 20170265 (2018).
11. Eikema, K. S., Walz, J. & Haensch, T. W. Continuous wave coherent Lyman-α radiation. *Phys. Rev. Lett.* **83**, 3828–3831 (1999).
12. Michan, J. M., Polovy, G., Madison, K. W., Fujiwara, M. C. & Momose, T. Narrowband solid state VUV coherent source for laser cooling of antihydrogen. *Hyperf. Interact.* **235**, 29–36 (2015).
13. Ahmadi, M. et al. Enhanced control and reproducibility of non-neutral plasmas. *Phys. Rev. Lett.* **120**, 025001 (2018).
14. Ahmadi, M. et al. Antihydrogen accumulation for fundamental symmetry tests. *Nat. Commun.* **8**, 681 (2017).
15. Maury, S. The antiproton decelerator: AD. *Hyperf. Interact.* **109**, 43–52 (1997).
16. Murphy, T. J. & Surko, C. M. Positron trapping in an electrostatic well by inelastic collisions with nitrogen molecules. *Phys. Rev. A* **46**, 5696–5705 (1992).
17. Surko, C. M., Greaves, R. G. & Charlton, M. Stored positrons for antihydrogen production. *Hyperf. Interact.* **109**, 181–188 (1997).
18. Andresen, G. B. et al. Trapped antihydrogen. *Nature* **468**, 673–676 (2010).
19. Andresen, G. B. et al. Evaporative cooling of antiprotons to cryogenic temperatures. *Phys. Rev. Lett.* **105**, 013003 (2010).
20. Hoecker, A. et al. TMVA—toolkit for multivariate data analysis. Preprint at https://arxiv.org/abs/physics/0703039 (2007).
21. Amole, C. et al. In situ electromagnetic field diagnostics with an electron plasma in a Penning-Malmberg trap. *New J. Phys.* **16**, 013037 (2014).
22. Andresen, G. B. et al. Confinement of antihydrogen for 1000 seconds. *Nat. Phys.* **7**, 558–564 (2011).
23. Smorra, C. et al. A parts-per-billion measurement of the antiproton magnetic moment. *Nature* **550**, 371–374 (2017).
24. Kuroda, N. et al. A source of antihydrogen for in-flight hyperfine spectroscopy. *Nat. Commun.* **5**, 3089 (2014).
25. Gabrielse, G. et al. Trapped antihydrogen in its ground state. *Phys. Rev. Lett.* **108**, 113002 (2012).
26. Doser, M. et al. Exploring the WEP with a pulsed cold beam of antihydrogen. *Class. Quantum Grav.* **29**, 184009 (2012).
27. Pérez, P. et al. The GBAR antimatter gravity experiment. *Hyperf. Interact.* **233**, 21–27 (2015).
28. Parthey, G. et al. Improved measurement on the hydrogen 1S–2S spectroscopy. *Phys. Rev. Lett.* **107**, 203001 (2011).
29. Beyer, A. et al. The Rydberg constant and proton size from atomic hydrogen. *Science* **358**, 79–85 (2017).
30. Fleurbaey, H. et al. New measurement of the 1S–3S transition frequency of hydrogen: contribution to the proton charge radius puzzle. *Phys. Rev. Lett.* **120**, 183001 (2018).
31. Mohr, P. J., Newell, D. B. & Taylor, B. N. CODATA recommended values of the fundamental physical constants: 2014. *Rev. Mod. Phys.* **88**, 035009 (2016).
32. Hamilton, P. et al. Antimatter interferometry for gravity measurements. *Phys. Rev. Lett.* **112**, 121102 (2014).

Acknowledgements All authors are members of the ALPHA Collaboration. This work was supported by: the European Research Council through its Advanced Grant programme (to J.S.H.); CNPq, FAPERJ, RENAFAE (Brazil); NSERC, CFI, NRC/TRIUMF, EHPDS/EHDRS (Canada); FNU (Nice Centre), Carlsberg Foundation (Denmark); ISF (Israel); STFC, EPSRC, the Royal Society and the Leverhulme Trust (UK); DOE, NSF (USA); and VR (Sweden). We are grateful for the efforts of the CERN Antiproton Decelerator team, without which these experiments could not have taken place. We thank P. Djuricanin (University of British Columbia) for his extensive help with the laser system. We thank J. Tonoli (CERN) and his staff as well as T. Mittertreiner (University of British Columbia) and his staff for extensive, time-critical help with machining and electrical works. We thank the staff of the Superconducting Magnet Division at Brookhaven National Laboratory for collaboration and fabrication of the trapping magnets. We thank C. Marshall (TRIUMF) for work on the ALPHA-2 cryostat. We thank D. Tommasini and A Milanese (CERN) for the fabrication of conventional magnets for ALPHA-2. We thank F. Besenbacher (Aarhus University) for timely support in procuring the ALPHA-2 external solenoid. We thank T. Miller (Ohio) for advice on the initial development of the pulsed laser system.

Author contributions This experiment was based on data collected using the ALPHA-2 antihydrogen trapping apparatus, designed and constructed by the ALPHA Collaboration using methods developed by the entire collaboration. The entire collaboration participated in the operation of the apparatus and the data-taking activities. Pulsed Lyman-α spectroscopy in ALPHA was proposed by M.C.F. and developed further by T.M., F.R., J.M.M., R.C., A.E., A.K. and M.C.F. The original laser was designed by T.M. and tested by J.M.M. and T.M. The laser system at CERN was implemented, commissioned and operated by J.M.M., R.C., A.E., A.K. and T.M. F.R. developed the simulation programme for laser interaction with magnetically trapped atoms. Electron cyclotron resonance magnetometry was developed by T.F., M.E.H. and W.N.H. Detailed analysis of the antiproton annihilation detector data was done by J.T.K.M. The manuscript was written by T.M., M.C.F. and J.S.H., with help from R.C., A.E., A.K., J.T.K.M., F.R. and A.C. The manuscript was then edited and improved by the entire collaboration.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0435-1.

Reprints and permissions information is available at http://www.nature.com/reprints.

Correspondence and requests for materials should be addressed to M.C.F. or J.S.H. or T.M.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

METHODS

Laser system for 121.6-nm light. The schematic diagram of the laser system is shown in Extended Data Fig. 1. The vacuum ultraviolet radiation at 121.6 nm is produced in two steps: frequency doubling of 730-nm pulses followed by third harmonic generation in a high-pressure gas cell[12]. The 730-nm pulses are produced by first seeding two titanium sapphire crystals with narrow line width ($<$100 kHz), continuous-wave radiation at 730 nm from a Toptica diode laser. The crystals are pumped by nanosecond pulses of the second harmonic of a Nd:YAG (neodymium-doped yttrium aluminium garnet) laser (Spectra Physics, pulse energy 300 mJ at 532 nm, with 10-Hz repetition rate). The generated 730-nm pulses have a pulse length and line width of 30 ns and 25 MHz, respectively. The 730-nm pulses are then converted to 365 nm by frequency doubling in a beta barium borate (BBO) crystal and then directed to the experimental zone. The third harmonic of 365 nm is generated in a high-pressure Kr/Ar gas cell (total pressure about 4 bar) after focusing the 365-nm pulses with an ultraviolet-grade lens (focal length 150 mm). The conversion efficiency from 365 nm to 121.6 nm is typically 10^{-6} to 10^{-7}. The generated 121.6-nm pulses are collimated by a lens of 90-mm focal length and vacuum-ultraviolet-grade and then directed to the trap region by three vacuum-ultraviolet-grade mirrors placed in the vacuum system used to steer the light towards the antihydrogen trap.

Laser pulses enter the atom trap ultrahigh vacuum through an MgF_2 window, which transmits about 65% of the incident light. The laser pulses cross the trap at an angle of about 2° to the direction of the trap axis, as shown in Fig. 1. Pulses exiting the trap are detected by a calibrated, solar-blind photomultiplier tube (Hamamatsu) to monitor the pulse energy and timing. The typical pulse width and line width at 121.6 nm are 12 ns and 65 MHz (FWHM, estimated). Each pulse has a pulse energy of 0.53–0.65 nJ in the trap region, and the repetition rate is 10 Hz, corresponding to an average power of 5–7 nW, and a peak power of 50–70 mW. The linear polarization is perpendicular to the direction of the magnetic field in order to drive the transition to the $2P_c$ state. The seed laser frequency is stabilized to a few megahertz by locking the laser to a wavemeter (HighFinesse). Before and after each experimental sequence, the frequency and line width of the 730-nm pulses were monitored by a custom-made, Fabry–Perot spectrometer having a free spectral range of 745 MHz, and a finesse of about 75.

Suppression of cosmic ray background. To detect annihilation events in the two observation windows—(1) 15.6-s magnet ramp-down to release remaining antiatoms and (2) 1-ms observation window after each laser pulse—we require different levels of background suppression. We optimized significance using an MVA-based, machine-learning algorithm to suppress unwanted background counts based on estimated populations of atoms in the apparatus per sequences.

The selection of variables has been described elsewhere[7]. The algorithm was trained and tested using a sample of 393,920 and 3,375,877 events for signal and background, respectively. The signal events were selected from periods with high rates of antihydrogen production in the apparatus, and thus contain less than 0.1% background. The background events were collected during periods when no antiprotons were present in the apparatus.

The trigger is configured to fire on the n-side of the ALPHA silicon hybrid[33]. The total configuration for the silicon vertex detector trigger requires more than two hits on the inner layer with one hit on the middle and outer layer.

The 15.6-s observation window. A classifier cut was chosen to optimize the significance for an expected 150 counts of signal and 320 counts of background (motivated to optimize significance of the signal when the laser is tuned to resonance in two sequences). The analysis gives a background rate of $0.217 \pm 0.006\ \mathrm{s}^{-1}$ and an efficiency of 0.851 ± 0.002 (statistical error only, one standard deviation) annihilations per detector trigger.

The 1-ms (post-laser pulse) observation window. A classifier cut was chosen to optimize the significance for 31 expected counts of signal and 243 counts of background in the range ±10 cm along the beam axis of the trap centre. The analysis gives a background rate of $0.049 \pm 0.003\ \mathrm{s}^{-1}$ and an efficiency of 0.807 ± 0.002 (statistical error only) annihilations per detector trigger. For the sample of 966 annihilation events reported here, the expected background is 14 events.

Simulation of laser interaction with trapped antihydrogen atoms. The trapped antihydrogen motion is simulated as described elsewhere[34] using the $\boldsymbol{B}$-field obtained from an interpolation of a Biot–Savart model of the ALPHA-2 mirror and octupole coils. The atoms are randomly distributed between the $1S_c$ and $1S_d$ states, which, for the purposes of the centre of mass motion, are treated as having the same magnetic moment, equal to that of an electron. To mimic the experiment, the simulated atoms are launched in the $n = 30$ state and allowed to radiatively cascade to the ground state, or are directly launched in the ground state to obtain two sets of initial conditions. The atoms are propagated for a few seconds before the laser is turned on, after which a laser pulse interacts with the atoms every 100 ms. At each laser pulse, the distance of the antihydrogen from the laser is computed and the laser intensity at that position is used to compute the excitation probability. The probability of the excited state decaying into the $1S_a$ or $1S_b$ states is also computed. A random number is compared to the product of these probabilities to determine whether the spin of the simulated atom flips for that laser pulse. If a flip occurs, the sign of the magnetic moment is reversed to give a high-field-seeking atom. The simulation is stopped when the position of the atom is outside the inner edge of the trap electrodes.

The statistics on properties like the Doppler width and the energy shift due to the Zeeman effect are accounted for by using the properties of the system at each laser pulse. The shift in frequency due to the Doppler effect is calculated from the velocity $\boldsymbol{v}$ at the time of the laser pulse:

$$\delta\omega = \frac{-\omega(\boldsymbol{v}\hat{k})}{c}$$

where $\hat{k}$ is the direction of propagation of the laser beam. The Zeeman shift for that laser shot is computed from the strength of the $\boldsymbol{B}$-field at the position of the antihydrogen. The $1S_c$ and $1S_d$ energies as a function of $\boldsymbol{B}$ are calculated as described in ref. [34], but the 2P energies are calculated using slightly more accurate equations which account for the difference between the positron magnetic moment and the Bohr magneton. The hyperfine splitting of the 2P states is at the level of a few tens of megahertz and, although included in the simulation, do not affect the results at the level presented here. The laser line width is included in the simulation by randomly shifting the laser frequency with a Gaussian distribution given by a FWHM of $2\pi \times 65$ MHz. These shifts are added to the detuning of the laser to obtain a total energy shift for that laser pulse of $\hbar\delta\omega$ relative to the line centre.

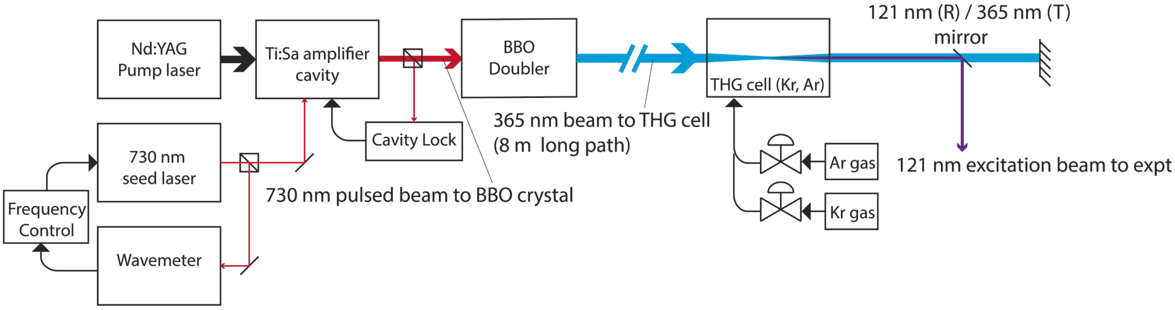
The probability that an excitation occurs during a laser pulse is proportional to the laser intensity at the position of the antihydrogen. The simulation assumes linearly polarized photons perpendicular to the direction of laser propagation, which is nearly parallel to $\boldsymbol{B}$. Thus, the polarization is nearly perpendicular to $\boldsymbol{B}$, and approximately half of the laser intensity can drive the transition. The transition dipole matrix element is equal to that of the $m = 0$ to $m = 1$ transition in hydrogen, multiplied by $\sin(\sigma)$ for the $2P_c$ level, where σ is a mixing angle; see equations (13) and (15) in ref. [34]. The transition probability is multiplied by the Lorentzian factor from the natural line width of the 2P states:

$$\frac{(\tilde{A}/2)^2}{(\delta\omega)^2 + (\tilde{A}/2)^2}$$

where $\tilde{A} = 2\pi \times 99.6$ MHz is the natural line width of the 2P state, and $\delta\omega$ is calculated as described in the previous paragraph. The probability of a spin flip in the decay is proportional to the probability in the 2P wavefunction with flipped spin: $\cos^2(\sigma)$ for the $2P_c$ state.

Data availability. The datasets generated and analysed during this study are available from J.S.H. on reasonable request.

33. Andresen, G. B. et al. The ALPHA detector: module production and assembly. *J. Instrum.* **7**, C01051 (2011).
34. Rasmussen, C. Ø., Madsen, N. & Robicheaux, F. Aspects of 1S–2S spectroscopy of trapped antihydrogen atoms. *J. Phys. B* **50**, 184002 (2017).


Nd:YAG
Pump laser
Ti:Sa amplifier
cavity
BBO
Doubler
121 nm (R) / 365 nm (T)
mirror
THG cell (Kr, Ar)
365 nm beam to THG cell
(8 m long path)
Cavity Lock
730 nm
seed laser
Frequency
Control
Wavemeter
730 nm pulsed beam to BBO crystal
Ar gas
Kr gas
121 nm excitation beam to expt

Extended Data Fig. 1 | Laser system. The figure shows a schematic of the 121.6-nm laser system for driving the 1S–2P transition. See Methods for details.

https://doi.org/10.1038/s41586-018-0480-9

Hundred-fold enhancement in far-field radiative heat transfer over the blackbody limit

Dakotah Thompson[1], Linxiao Zhu[1], Rohith Mittapally[1], Seid Sadat[1], Zhen Xing[2], Patrick McArdle[2], M. Mumtaz Qazilbash[2], Pramod Reddy[1,3]* & Edgar Meyhofer[1]*

Radiative heat transfer (RHT) has a central role in entropy generation and energy transfer at length scales ranging from nanometres to light years[1]. The blackbody limit[2], as established in Max Planck's theory of RHT, provides a convenient metric for quantifying rates of RHT because it represents the maximum possible rate of RHT between macroscopic objects in the far field—that is, at separations greater than Wien's wavelength[3]. Recent experimental work has verified the feasibility of overcoming the blackbody limit in the near field[4–7], but heat-transfer rates exceeding the blackbody limit have not previously been demonstrated in the far field. Here we use custom-fabricated calorimetric nanostructures with embedded thermometers to show that RHT between planar membranes with sub-wavelength dimensions can exceed the blackbody limit in the far field by more than two orders of magnitude. The heat-transfer rates that we observe are in good agreement with calculations based on fluctuational electrodynamics. These findings may be directly relevant to various fields, such as energy conversion, atmospheric sciences and astrophysics, in which RHT is important.

RHT has a key role in diverse processes, from controlling the temperature of planetary atmospheres[1] and planetary formation[8] to radiative cooling[9] and thermophotovoltaic energy conversion[10]. Recent studies of near-field RHT[4–7] have experimentally shown that RHT between surfaces separated by nanoscale gaps (gaps below the dominant thermal wavelength of $\lambda_{\mathrm{th}} \approx 10\,\mu\mathrm{m}$ at room temperature) can greatly exceed the fundamental blackbody limit[2] because of evanescent modes that are present in the near field. Such near-field effects are now being assessed extensively for their potential in energy conversion and thermal management technologies[11,12]. However the feasibility of enhancing RHT substantially in the far field (that is, the region where no evanescent modes are present) when the surfaces are separated by gaps larger than λ_{th} remains unexplored experimentally.

Planck's far-field RHT (FF-RHT) theory[2], which is based on a 'ray' picture of photon transport, predicts that given two finite-sized objects of surface areas A_1 and A_2 (Fig. 1a), the maximum possible conductance ($G_{\max}$) between these objects in the blackbody limit is given by $G_{\max} = 4\sigma T^3 A_1 F_{12} = 4\sigma T^3 A_2 F_{21}$, where F_{12} and F_{21} are geometric view factors that satisfy the reciprocity relationship $A_1 F_{12} = A_2 F_{21}$ (see Supplementary Fig. 1 or ref. [3]), σ is the Stefan–Boltzmann constant and T is the absolute temperature. Planck himself pointed out that the blackbody limit is not applicable when (1) the size of the gap between the objects is less than λ_{th} or (2) one or more dimensions of the objects are smaller than λ_{th}. As briefly described above, recent experimental measurements[5] of radiative thermal conductances between planar macroscopic surfaces (or objects) separated by gap sizes below λ_{th} showed extraordinary increases in comparison to the expected blackbody limit. Therefore, one may expect that the blackbody limit may also be readily overcome in the far field for objects of sub-wavelength (less than λ_{th}) thickness; for example, for sub-wavelength-diameter spherical particles, which feature sharp resonances in their absorption cross-section. However, both our detailed computational analysis (see Supplementary Fig. 2) and past computations[13–15] suggest that such enhancements are not anticipated for metallic spheres, and very small increases (by a factor of a few) may be expected for dielectric spheres or metallic cylinders. By strong contrast, interesting analytical work[16] performed recently indicated that for sub-wavelength objects, there may be no limit to how much thermal emission and absorption can exceed the blackbody limit.

Experimental work[17–19] on individual sub-wavelength structures has been largely limited to probing emission spectra because quantifying heat output or performing RHT measurements between these structures is very challenging. Notably, a recent experiment[19] that probed emissions from sub-wavelength cylindrical $\mathrm{SiO_2}$ wires revealed that the net radiative rates from these cylinders are actually much smaller than those predicted by the blackbody limit. Taken together, the question of whether RHT between sub-wavelength structures in the far field can greatly exceed Planck's blackbody limit remains unanswered and has never been addressed experimentally.

To explore whether very large enhancements over the blackbody limit can be accomplished we chose to work with thin planar structures (similar to those shown at the bottom of Fig. 1a), where one of the dimensions (the thickness, t) was chosen to be much smaller than λ_{th} at room temperature. The planar device geometries selected in this work represent shapes often employed in micro- and nano-electromechanical systems and nanoscale conduction studies[20,21] and are experimentally tractable for quantitative RHT measurements. Specifically, we custom-fabricated a series of devices featuring two coplanar silicon nitride (SiN) membranes (Fig. 1c) with thickness between 270 nm and about 11 μm (see Methods and Extended Data Fig. 1 for information regarding device fabrication). Each of these membranes is suspended by thin long beams and features an integrated serpentine Pt resistor (see inset of Fig. 1c and Fig. 1d) that can be employed to either heat a membrane or measure its temperature rise. Further, the separation between the membranes was chosen to be 20 μm to ensure that near-field effects are negligible, because $\lambda_{\mathrm{th}} \approx 10\,\mu\mathrm{m}$ at 300 K (see Supplementary Fig. 15, where we computationally show that near-field effects are less than 0.1% at 300 K). Our devices feature internal tensile stresses that ensure that the membranes have excellent planarity and are coplanar, as confirmed by laser scanning confocal microscopy; see Fig. 1e, f (additional details in Methods and Extended Data Fig. 2).

The basic strategy to quantify the RHT between the planar device structures in our measurements is to increase the temperature of one of the membranes by Joule heating via the integrated Pt heater, and then measure the temperature rise in the unheated membrane due to RHT from the hot to the cold membrane (see Fig. 1d). To perform the desired measurements, the devices were first placed in a vacuum of less than 10^{-6} torr (at room temperature), which attenuated parasitic heat conduction between the membranes through residual air molecules, to negligible levels (see Methods for quantification). Next, one of the planar devices, the emitter, was heated by a sinusoidal current with

---

[1]Department of Mechanical Engineering, University of Michigan, Ann Arbor, MI, USA. [2]Department of Physics, College of William and Mary, Williamsburg, VA, USA. [3]Department of Materials Science and Engineering, University of Michigan, Ann Arbor, MI, USA. *e-mail: pramodr@umich.edu; meyhofer@umich.edu

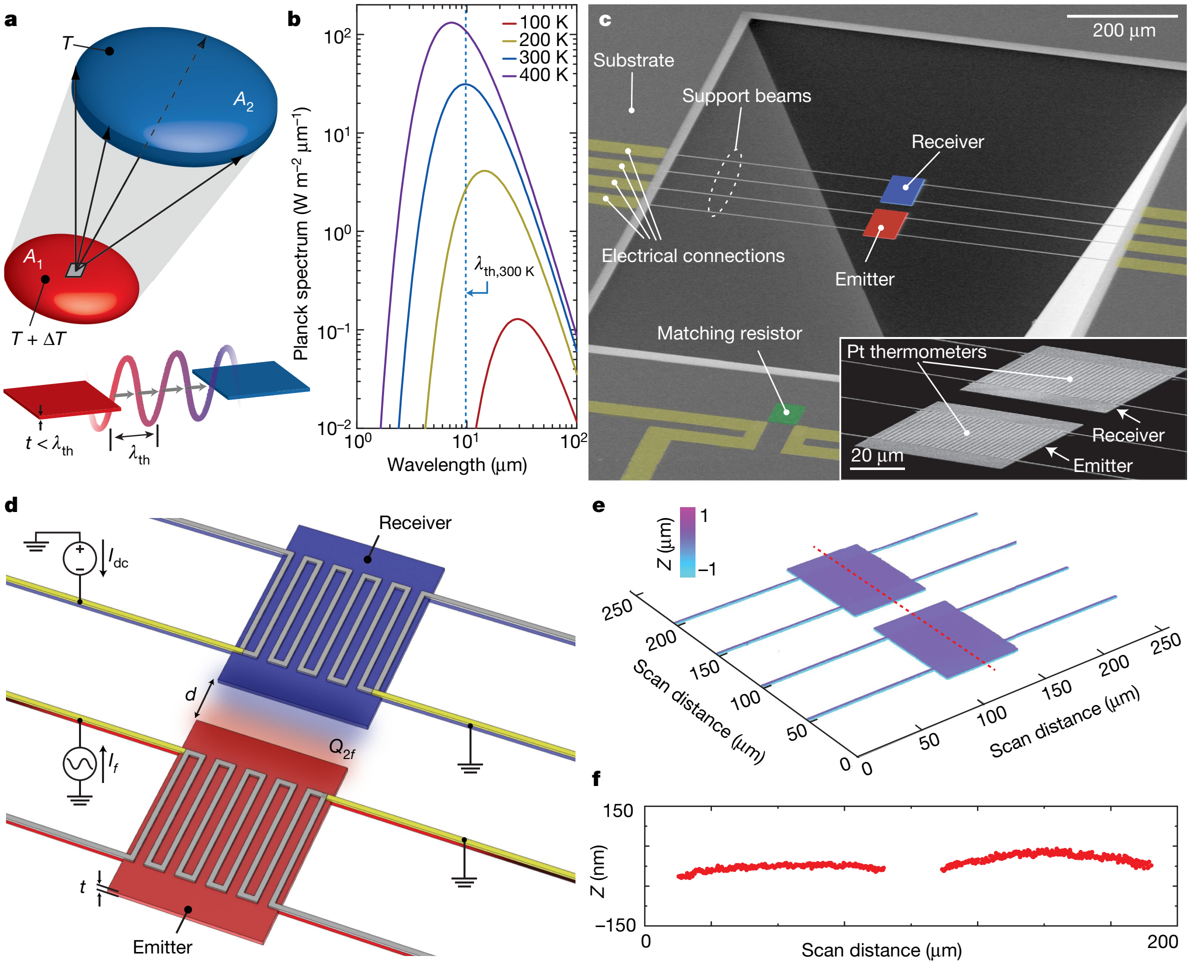


Fig. 1 | Probing RHT between sub-wavelength structures. a, Schematic describing the heat transfer between finite-sized objects of surface areas A_1 and A_2 and sub-wavelength plates or membranes. Here λ_{th} refers to Wien's wavelength, which is about 10 μm at room temperature (see **b**). Red indicates higher temperature ($T + \Delta T$) and blue indicates lower temperature (T). **b**, Spectral emissive power for blackbodies at various temperatures. **c**, Nanofabricated experimental platform used to probe RHT. The receiver and emitter devices, made from SiN, are suspended by support beams from a substrate and feature embedded Pt resistance thermometers–heaters (see Methods). Further, a matching resistor is integrated into the substrate. **d**, Schematic representation of the nanofabricated devices. One of the membranes, the emitter, is heated by passing a sinusoidal current of amplitude I_f and frequency f via the integrated Pt heater, which results in a heat current from the heater to the receiver membrane. The heat current (Q_{2f}) is measured by monitoring the temperature rise of the receiver via the integrated Pt resistance thermometer (using a d.c. current I_{dc}). The thickness of the membranes and the gap size between them are represented by t and d, respectively. **e**, Scans of 2-μm-thick nanofabricated devices obtained with a confocal microscope reveal that the membranes have excellent planarity. **f**, Profile along the red dashed line in **e**, showing that any deviations from planarity (Z) are smaller than 100 nm.

amplitude I_f and frequency f, resulting in temperature oscillations ($\Delta T_{\mathrm{emit},2f}$) at $2f$, which were measured following protocols that we developed previously[22,23] (see Methods for a description). The radiative heat flux (Q_{2f}) from the emitter to the second device, called the receiver, resulted in temperature oscillations of the receiver's membrane ($\Delta T_{\mathrm{rec},2f}$), which we measured by supplying a d.c. current, I_{dc}, to the Pt resistor and measuring the voltage fluctuations at $2f$ (see Methods). Results from measurements performed on a 486-nm-thick device are shown in Fig. 2a, where $\Delta T_{\mathrm{rec},2f}$ is plotted as a function of $\Delta T_{\mathrm{emit},2f}$. The inset of Fig. 2a shows the expected temperature field of the emitter, estimated from finite-element simulations. It can be seen that the suspended membrane of the emitter has a uniform temperature, whereas there is a linear temperature drop in the beams.

The effective radiative conductance ($G_{\mathrm{rad,eff}}$) between the emitter and receiver is estimated from $G_{\mathrm{rad,eff}} = G_{\mathrm{beams,rec}} \times [(\Delta T_{\mathrm{emit}}/\Delta T_{\mathrm{rec}}) - 1]^{-1}$, where $G_{\mathrm{beams,rec}}$ is the experimentally determined thermal conductance of the receiver's beams (see Extended Data Fig. 3). The measured $G_{\mathrm{rad,eff}}$ for devices of various thicknesses is shown in Fig. 2b (black circles). It can be seen that $G_{\mathrm{rad,eff}}$ decreases monotonically with decreasing thickness, as one would expect from basic FF-RHT theory[2]. To quantitatively compare the experimental data with the predictions of FF-RHT theory, we computed $G_{\mathrm{rad,eff}}$ by performing comprehensive RHT simulations using a finite-element method (see Supplementary Fig. 3 for the details of the simulations), which fully accounts for the geometric view factors[3] and assumes the emissivity of the devices to be 1 (that is, the blackbody limit). The measured $G_{\mathrm{rad,eff}}$ was found to be much larger than the simulated blackbody value (black triangles in Fig. 2b) for the thinner devices. Specifically, for membranes with a thickness of 270 nm, the ratio of the measured radiative conductance to the simulated blackbody conductance (red circles) is more than 100. However, for the thickest device (about 11 μm thick) the predictions are in good agreement with the experimental data. The data clearly show that—for sub-wavelength ($t \ll \lambda_{\mathrm{th}}$) planar structures—the blackbody limit can be readily overcome. In fact, such large enhancements over the blackbody limit have only been reported in the near field[5,24], where gap sizes have been reduced to tens of nanometres, and never in the far field. We note that we performed several control experiments (see Methods) that unambiguously ruled out the possibility that the observed $G_{\mathrm{rad,eff}}$ was due to other heat-transport mechanisms (such as conduction through air or the substrate, or spurious electrical signals due to capacitive coupling between the Pt heater and sensor) instead of RHT.

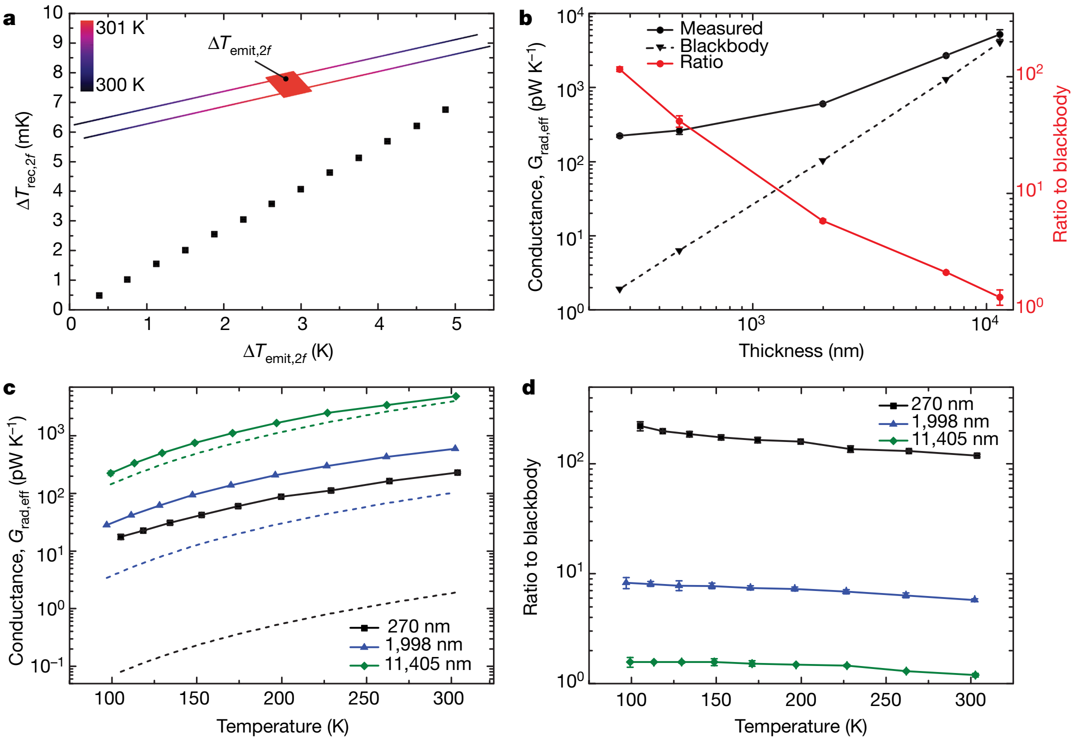


Fig. 2 | Radiative conductance between membranes with sub-wavelength dimensions. a, Measured temperature rise of the receiver ($\Delta T_{rec,2f}$) due to radiative heat flow from a heated emitter. Error bars (1σ, smaller than symbols) were obtained from the standard deviation of measurements performed over a 50-min-long period (see Methods for details). The inset shows the temperature profile of the emitter and the beams from finite-element simulations. **b**, Measured effective radiative conductance $G_{rad,eff}$ at room temperature (black circles; left axis) for membrane thicknesses from about 11 μm to 270 nm. Errors for $G_{rad,eff}$ (1σ, usually smaller than the symbols) were determined from regression analysis of the data shown in **a** (see Methods for details). The computed $G_{rad,eff}$ (black triangles; left axis; from FF-RHT theory, with an emissivity of 1 for all objects) is considerably smaller for thin devices and the ratio of the measured to the simulated radiative conductance increases steeply with reduced devices thicknesses (red circles; right axis). **c**, Measured $G_{rad,eff}$ as a function of temperature and device thickness. Simulated results (dashed lines; from FF-RHT theory) are found to substantially under-predict $G_{rad,eff}$ at all temperatures, particularly for the thinner devices. Similar data for 486-nm-thick and 6,712-nm-thick devices are shown in Extended Data Fig. 5a. **d**, Ratio of the measured to the estimated (using FF-RHT theory) conductance as a function of temperature. For the thinner devices the measured conductance exceeds predictions by a factor of more than 100. Error bars in **c** and **d** were obtained following the same approach as for **b**.

To further understand the effect of λ_{th} on the RHT properties of our devices, we performed measurements in which the temperature of the devices was progressively decreased from 300 K to 100 K. Because FF-RHT theory[2] predicts that $\lambda_{th}T = 2{,}898\,\mu\mathrm{m\,K}$ (Wien's law), the thermal wavelength increases with decreasing temperature, resulting in an even larger difference between λ_{th} and the thickness of the devices. The data obtained from these measurements for three different thicknesses (270 nm, 1,998 nm and 11,405 nm) are shown in Fig. 2c (data for other thicknesses are presented in the Methods section), which demonstrates that the measured thermal conductance declines monotonically with decreasing temperature. The corresponding simulated FF-RHT thermal conductances (dashed lines) are found to be much smaller than the measured conductances for the thinner devices. The ratios of the measured to the simulated conductance for the three devices are shown in Fig. 2d. It is clear from these data that the disagreement between the measured and computed conductances persists at lower temperatures. In fact, for the thickest device the agreement between measurements and simulation is very good at 300 K, but at lower temperatures the measured conductances are larger than those predicted by simulations—again highlighting the failure of FF-RHT theory when $t \ll \lambda_{th}$. The contribution of near-field effects is negligible even at 100 K (less than 2%; see Supplementary Fig. 15). Taken together, these data raise the important question of whether the observed 100-fold enhancement in comparison to the blackbody limit can be explained quantitatively.

To address this question we resorted to the framework of fluctuational electrodynamics[25,26] (FED)—a theory that has been extensively employed in the past to study near-field RHT because it does not rely on the assumptions of FF-RHT theory[2]. In our simulations we implemented our FED model using the boundary-element method[27,28] (BEM). BEM-based FED models, which have been developed recently, enable an exact analysis of radiation for complex geometries such as our membranes and the support beams (see Supplementary Fig. 5 for details). However, the Pt lines were not incorporated into our FED models to keep the computations tractable. For the FED calculations, our device geometries were discretized into surface elements, as shown, for example, in Fig. 3a (details of meshing and convergence criteria are discussed in Supplementary Fig. 6) for the BEM mesh employed in simulations of our 2-μm-thick device. Key to performing the desired calculations are the dielectric properties of SiN, which have been previously characterized[29] but were also measured for our films (see Supplementary Fig. 4).

The calculated $G_{rad,eff}$ from our FED simulations is shown in Fig. 3b and—in contrast to FF-RHT-theory-based simulations—are in excellent agreement with the measured $G_{rad,eff}$ for all thicknesses (all deviations are lower than 30%). To obtain deeper insight into why thinner devices support much larger conductances than predicted by FF-RHT theory, we computed the spectral heat conductance for each device and normalized it by the thickness of the device. The normalized data, shown in Fig. 3c (for three thicknesses), are found to be much larger for the thinnest device than for thicker devices. These deviations arise primarily at frequencies where there is a higher density of guided modes that are confined to the SiN membranes, as revealed by the dispersion relation for guided modes in 270-nm-thick SiN membranes (see inset of Fig. 3c and Supplementary Fig. 16 for additional details).

Finally, we performed FED-based calculations of the effective conductance of the devices as a function of device temperature. For

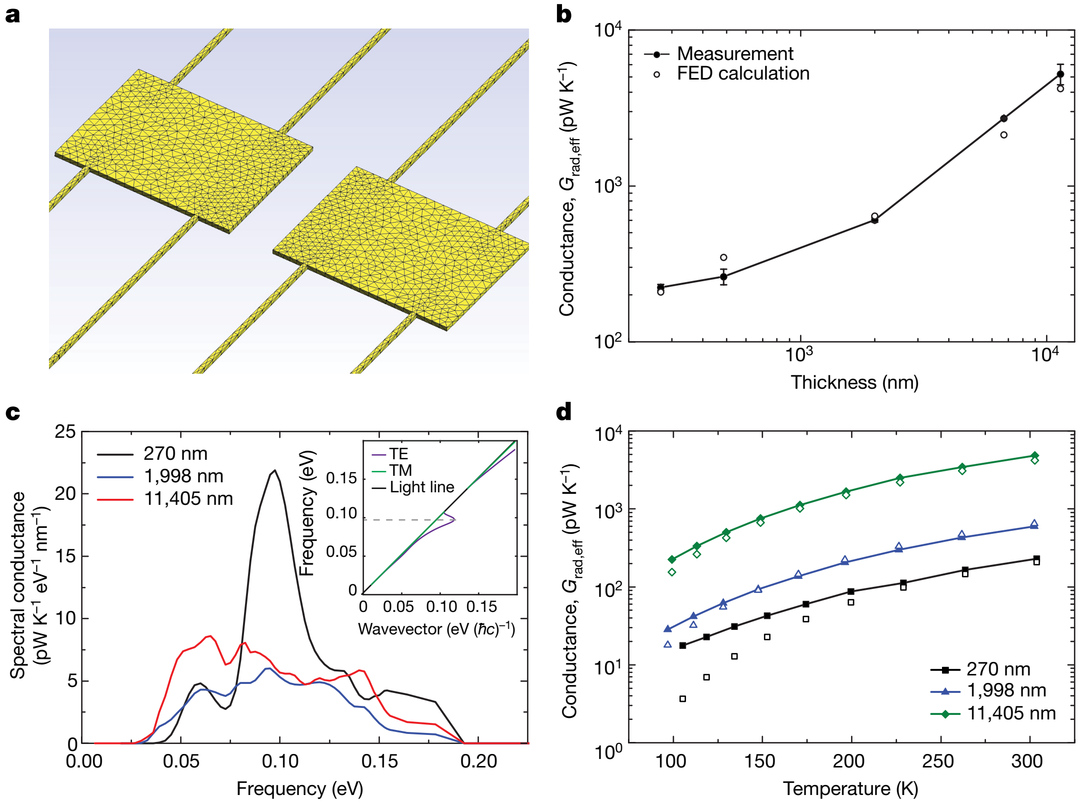


Fig. 3 | Comparison of experimental data with FED-based simulations. **a**, Schematic depicting the meshing scheme employed in our BEM calculations. The exact geometry of the membranes is taken into account. However, the metal lines are excluded to make the computations tractable. **b**, Comparison of the measured radiative conductance (solid black circles) with the calculated radiative conductance obtained from FED simulations (black open circles) at room temperature. Error bars were obtained in a similar fashion to those in Fig. 2b. **c**, Spectral heat conductance, normalized by the thickness of the devices, for devices of various thicknesses at room temperature. The spectral heat transfer is dominated by a narrow band of frequencies for the thinner devices, which corresponds to the range over which the dispersion relation for guided modes becomes highly dispersive, indicating a larger density of states in a small range of frequencies. The inset shows the dispersion relation of guided modes (transverse electric (TE) and transverse magnetic (TM)) in the 270-nm-thick SiN membrane. The light line in vacuum is also shown (see Supplementary Fig. 16 for additional details). Similar data for the 486-nm- and 6,712-nm-thick devices are shown in Extended Data Fig. 5b. **d**, Measured (filled symbols) and computed (with FED simulations; open symbols) $G_{\mathrm{rad,eff}}$ at various temperatures agree well with each other at all temperatures, except for some discrepancies at the lowest temperatures (see Supplementary Figs. 13 and 14 for a discussion about the possible causes of these discrepancies). Error bars (obtained in a similar fashion to those in Fig. 2c; see Methods) for the measured data are smaller than the symbols.

all the devices except the thinnest one, the FED-based computations are found to agree very well with the measured results over all temperatures. For the thinnest device the agreement between FED-based calculations and measurements is poor at lower temperatures (100–200 K). Although this disagreement was initially puzzling, it can be largely resolved when we consider the contributions from the metallic lines, which were ignored in our FED-based calculations. Specifically, our calculations for metallic lines (see Supplementary Figs. 13, 14), which are consistent with past work[14], show that at low temperatures the contributions of the metallic lines to $G_{\mathrm{rad,eff}}$ can be as large as several picowatts per kelvin and therefore can dominate the RHT between the thinner devices.

The above discussion shows that although the unique FF-RHT phenomena observed in this work cannot be explained with FF-RHT theory[3], they are consistent with the predictions of FED. The experimental data and analysis presented above raise the following questions: why does FF-RHT theory fail to explain the experimental data? How general are the observed deviations—that is, can similar enhancements be obtained from other materials and geometries? To address these questions we considered a simplified geometry where we analysed the RHT for isolated membranes similar in dimensions to those used in our experiments, but without beams. We first used the electromagnetic solver SCUFF-EM[30] to compute the magnitude of the time-averaged Poynting vector (S_{FED}) for membranes of various thicknesses as a function of the polar angle (θ) for a fixed azimuthal angle and distance (60 μm; see Fig. 4a) at room temperature. Next, we computed the heat fluxes (S_{bb}) at the same locations using FF-RHT theory and treating the membranes as blackbodies. To compare these two heat fluxes we show in Fig. 4a the heat flux enhancement ratio, HFER = $S_{\mathrm{FED}}/S_{\mathrm{bb}}$ (see Supplementary Fig. 7 for more details), for membranes of three thicknesses (results for membranes with other thicknesses are shown in Supplementary Fig. 10). It can be seen that for the thinner membranes, HFER peaks sharply in the in-plane direction, with a value that is more than 10 times larger than the blackbody limit. By contrast, the out-of-plane Poynting flux is attenuated substantially below the blackbody limit. However, this behaviour is not observed for the thicker structures, whose emission and absorption characteristics are very similar to those predicted by FF-RHT theory.

Next, we used SCUFF-EM to compute the absorption cross-sections, σ_{FED}, as a function of the polar angle, θ' (see Fig. 4a), and compared them to the absorption cross-sections, σ_{bb}, calculated from FF-RHT theory by treating the membranes as blackbodies (see Supplementary Fig. 8). The computed absorption cross-section enhancement ratio, ACER = $\sigma_{\mathrm{FED}}/\sigma_{\mathrm{bb}}$, is shown at the bottom of Fig. 4a. It can be seen that ACER has a similar polar-angle dependence to HFER. To evaluate whether these increases in HFER and ACER account for the enhanced heat-transfer rates, we also computed the heat-transfer rates (Q_{FED}) between adjacent membranes (20 μm gap) using FED and estimated its enhancement in comparison to the blackbody limit (Q_{bb}) for various membrane thicknesses. The ratio $Q_{\mathrm{FED}}/Q_{\mathrm{bb}}$ (Fig. 4b) is much larger for

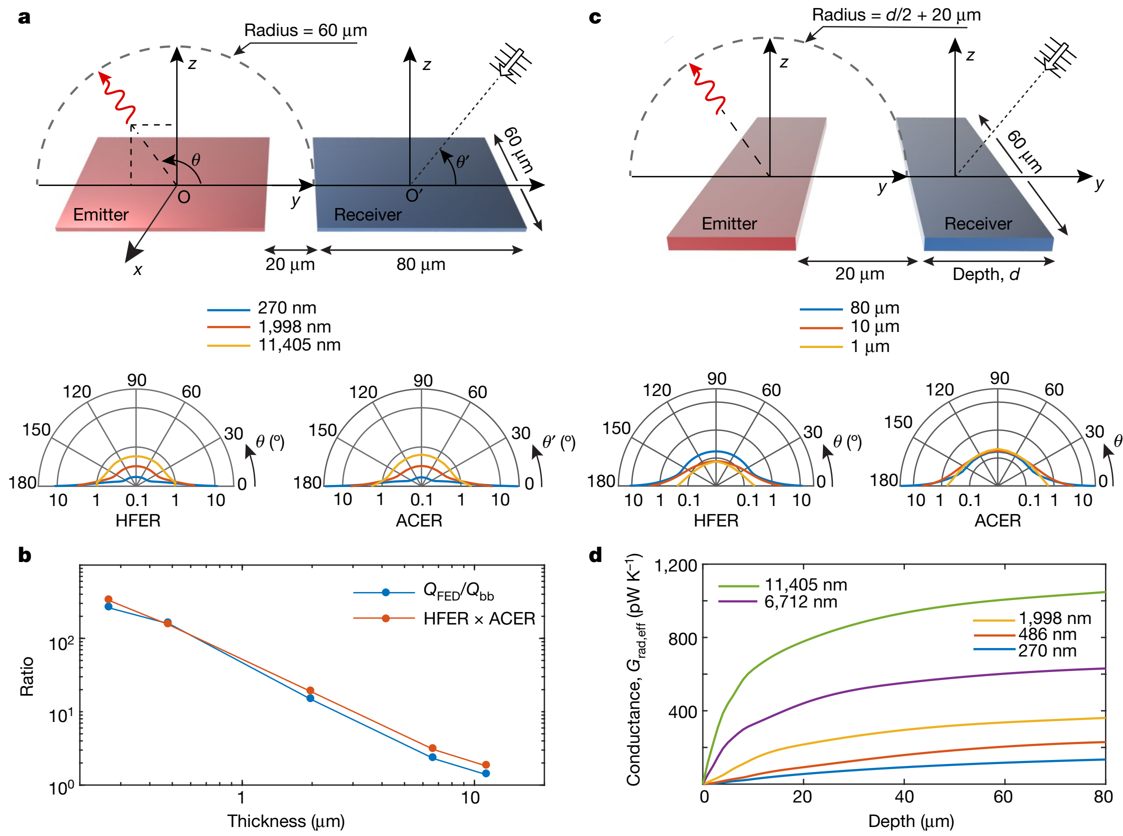


Fig. 4 | Analysis of the observed violations of the blackbody limit. **a**, Simplified geometry of membranes (excluding beams) used in the analysis of the emissions and absorption. The lateral dimensions are kept constant but the thickness of the membranes is varied in the calculations. The bottom graphs show the computed HFER as a function of polar angle, θ (left; at a fixed azimuthal angle and a distance of 60 µm from the origin O) and the computed angle-dependent ACER (right). It can be seen that HFER and ACER are considerably larger in the in-plane direction for the thinner devices. Additional calculation results for HFER at larger distances are shown in Supplementary Fig. 9. **b**, The ratio of the radiative thermal conductance computed using FED (Q_{FED}) to that obtained from FF-RHT theory (Q_{bb}) is shown along with the product of HFER and ACER. **c**, Computed HFER and ACER for 270-nm-thick membranes as a function of polar angle, for various depths (d) at a fixed distance of $d/2 + 20$ µm from the volumetric centre of the membrane. **d**, Computed radiative conductance for the configurations shown in **c** as a function of membrane depth. It can be seen that $G_{\mathrm{rad,eff}}$ decreases with decreasing depth.

thinner devices than for thicker ones (for which the ratio is about 1) and is in good agreement with the product of HFER and ACER. Therefore, we conclude that the experimentally observed enhancements of RHT can be explained by the enhanced HFER and ACER. Additional analysis on Au membranes (see Supplementary Figs. 11, 12) demonstrates that substantial enhancements can also be obtained by very thin metallic structures, for reasons similar to those described above. We also note that the 100-fold enhancements in the measured RHT relative to that predicted by FF-RHT theory persist as the gap sizes increase to the millimetre scale, as confirmed by additional computations and experiments presented in the Supplementary Information (see Supplementary Figs. 19 and 20).

Finally, to understand the dependence of the RHT on the depth of the membranes, we computationally analysed HFER and ACER for 270-nm-thick, 60-µm-wide membranes while the depth was progressively reduced (Fig. 4c). The results suggest that as the depth is decreased, HFER and ACER become progressively smaller. We performed FED-based calculations to understand the dependence of radiative conductance on the membrane depth. The results (Fig. 4d) show that the radiative conductance decreases rapidly with decreasing depth, highlighting that the depth has a critical role in the observed enhancements.

Our experiments provide direct evidence that the blackbody limit can be overcome by orders of magnitude in the far field. Our results represent a considerable advance over past experimental work, where enhancements compared to the blackbody limit were observed only in the near field. Further, our complementary computations provide additional insights into FF-RHT between nanoscale objects in terms of enhanced emissions and absorption cross-sections. This work highlights the potential of using nanoscale membranes for enhancing and controlling thermal absorption. Further work on optimizing and enhancing such absorption characteristics may have a dramatic impact on thermal energy capture for thermal to electric energy conversion. Finally, the insights obtained from this work may also bear upon the modelling of planetary atmospheres[31], proto-planetary formation[8] and astrophysical phenomena[32], where thermal radiation has an important role.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0480-9.

Received: 22 October 2017; Accepted: 26 June 2018;
Published online: 03 September 2018

1. Chandrasekhar, S. *Radiative Transfer* (Clarendon Press, Oxford, 1950).
2. Planck, M. *The Theory of Heat Radiation* [transl. Masius, M.] (P. Blakiston's Son & Co., Philadelphia, 1914).
3. Modest, M. F. *Radiative Heat Transfer* 3rd edn (Academic Press, New York, 2013).
4. Kim, K. et al. Radiative heat transfer in the extreme near field. *Nature* **528**, 387–391 (2015).
5. Song, B. et al. Radiative heat conductances between dielectric and metallic parallel plates with nanoscale gaps. *Nat. Nanotechnol.* **11**, 509–514 (2016).
6. Shen, S., Narayanaswamy, A. & Chen, G. Surface phonon polaritons mediated energy transfer between nanoscale gaps. *Nano Lett.* **9**, 2909–2913 (2009).
7. Rousseau, E. et al. Radiative heat transfer at the nanoscale. *Nat. Photon.* **3**, 514–517 (2009).

8. Reipurth, B., Jewitt, D. & Keil, K. (eds) *Protostars and Planets V* (Univ. Arizona Press, Tucson, 2007).
9. Raman, A. P., Anoma, M. A., Zhu, L., Rephaeli, E. & Fan, S. Passive radiative cooling below ambient air temperature under direct sunlight. *Nature* **515**, 540–544 (2014).
10. Guler, U., Boltasseva, A. & Shalaev, V. M. Refractory plasmonics. *Science* **344**, 263–264 (2014).
11. Ben-Abdallah, P. & Biehs, S.-A. Near-field thermal transistor. *Phys. Rev. Lett.* **112**, 044301 (2014).
12. Otey, C. R., Lau, W. T. & Fan, S. Thermal rectification through vacuum. *Phys. Rev. Lett.* **104**, 154301 (2010).
13. Fernández-Hurtado, V., Fernández-Domínguez, A. I., Feist, J., García-Vidal, F. J. & Cuevas, J. C. Super-Planckian far-field radiative heat transfer. *Phys. Rev. B* **97**, 045408 (2018).
14. Golyk, V. A., Krüger, M. & Kardar, M. Heat radiation from long cylindrical objects. *Phys. Rev. E* **85**, 046603 (2012).
15. Kattawar, G. & Eisner, M. Radiation from a homogeneous isothermal sphere. *Appl. Opt.* **9**, 2685–2690 (1970).
16. Biehs, S.-A. & Ben-Abdallah, P. Revisiting super-Planckian thermal emission in the far-field regime. *Phys. Rev. B* **93**, 165405 (2016).
17. Au, Y.-Y., Skulason, H. S., Ingvarsson, S., Klein, L. J. & Hamann, H. F. Thermal radiation spectra of individual subwavelength microheaters. *Phys. Rev. B* **78**, 085402 (2008).
18. Schuller, J. A., Taubner, T. & Brongersma, M. L. Optical antenna thermal emitters. *Nat. Photon.* **3**, 658–661 (2009).
19. Wuttke, C. & Rauschenbeutel, A. Thermalization via heat radiation of an individual object thinner than the thermal wavelength. *Phys. Rev. Lett.* **111**, 024301 (2013).
20. Hochbaum, A. I. et al. Enhanced thermoelectric performance of rough silicon nanowires. *Nature* **451**, 163–167 (2008).
21. Kim, P., Shi, L., Majumdar, A. & McEuen, P. L. Thermal transport measurements of individual multiwalled nanotubes. *Phys. Rev. Lett.* **87**, 215502 (2001).
22. Sadat, S. et al. Room temperature picowatt-resolution calorimetry. *Appl. Phys. Lett.* **99**, 043106 (2011).
23. Sadat, S., Meyhofer, E. & Reddy, P. Resistance thermometry-based picowatt-resolution heat-flow calorimeter. *Appl. Phys. Lett.* **102**, 163110 (2013).
24. St-Gelais, R., Zhu, L. X., Fan, S. H. & Lipson, M. Near-field radiative heat transfer between parallel structures in the deep subwavelength regime. *Nat. Nanotechnol.* **11**, 515–519 (2016).
25. Polder, D. & Van Hove, M. Theory of radiative heat transfer between closely spaced bodies. *Phys. Rev. B* **4**, 3303–3314 (1971).
26. Rytov, S. M., Kravtsov, I. A. & Tatarskii, V. I. *Principles of Statistical Radiophysics* 2nd edn (Springer, Berlin, 1987).
27. Otey, C. R., Zhu, L., Sandhu, S. & Fan, S. Fluctuational electrodynamics calculations of near-field heat transfer in non-planar geometries: a brief overview. *J. Quant. Spectrosc. Radiat. Transf.* **132**, 3–11 (2014).
28. Rodriguez, A. W., Reid, M. H. & Johnson, S. G. Fluctuating-surface-current formulation of radiative heat transfer for arbitrary geometries. *Phys. Rev. B* **86**, 220302 (2012).
29. Cataldo, G. et al. Infrared dielectric properties of low-stress silicon nitride. *Opt. Lett.* **37**, 4200–4202 (2012).
30. Reid, M. T. H. & Johnson, S. G. Efficient computation of power, force and torque in BEM scattering calculations. *IEEE Trans. Antenn. Propag.* **63**, 3588–3598 (2015).
31. Goody, R. M. & Yung, Y. L. *Atmospheric Radiation: Theoretical Basis* 2nd edn (Oxford Univ. Press, Oxford, 1989).
32. Draine, B. & Weingartner, J. C. Radiative torques on interstellar grains: I. Superthermal spinup. *Astrophys. J.* **470**, 551–565 (1996).

Acknowledgements P.R. and E.M. acknowledge support from the Office of Naval Research under award number N00014-16-1-2672 (device fabrication), from the Army Research Office under award number W911NF-18-1-0004 (computational modelling) and from Department of Energy Basic Energy Sciences under award number DE-SC0004871 (instrumentation for radiative measurements). M.M.Q. is grateful for support from the National Science Foundation under award number DMR-1255156 (characterization of dielectric functions). L.Z. thanks C. R. Otey for discussions. We acknowledge the Lurie Nanofabrication Facility for facilitating the fabrication of the devices and the Advanced Research Computing Technology Services at the University of Michigan for computational resources.

Reviewer information *Nature* thanks P. Bharadwaj and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions P.R. and E.M. conceived the work. D.T., R.M. and S.S. fabricated the devices and performed the RHT experiments. L.Z. performed the calculations. Z.X. and P.M. performed spectroscopic measurements of the dielectric functions of silicon nitride under the guidance of M.M.Q. The manuscript was written by D.T., L.Z., P.R. and E.M. with comments and input from all authors.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0480-9.

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0480-9.

Reprints and permissions information is available at http://www.nature.com/reprints.

Correspondence and requests for materials should be addressed to P.R. or E.M.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Fabrication of the devices. The fabrication process of the devices is shown in Extended Data Fig. 1. A low-stress silicon nitride (SiN) film with a nominal stoichiometry of Si_3N_4 and a thickness of 270 nm, 486 nm or 1,998 nm is first deposited onto a 500-μm-thick silicon wafer (Step 1) via low-pressure chemical vapour deposition (LPCVD). A 30-nm-thick Pt film heater–thermometer is then patterned onto the SiN layer using a lift-off process (Step 2). Then, 100-nm-thick gold electrical connections are patterned onto the SiN layer using a similar lift-off process (Step 3). The contour of the suspended membranes and support beams is formed by reactive ion etching through the SiN layer on the front side of the wafer (Step 4). Windows are created in the SiN layer on the back of the wafer by reactive ion etching (Step 5), followed by a potassium hydroxide (KOH) etch to create a through-hole in the silicon handle wafer and release the suspended membrane devices (Step 6). To create suspended devices with even greater thickness (6,712 nm and 11,405 nm), 2–4-μm-thick SiN layers were symmetrically deposited on both sides of the 1,998-nm-thick devices by plasma-enhanced chemical vapour deposition (PECVD) after the release step (Step 7). The thickness of the devices was measured using spectroscopic reflectometry (resolution of about 1–2 nm).

Scanning electron microscope and optical microscope images of one of the fabricated devices are shown in Extended Data Fig. 1b and c, respectively. The suspended membranes are 60 μm × 80 μm in size, and the support beams are 400 μm long and 2 μm wide (see Extended Data Fig. 1).

Characterization of the flatness and coplanarity of adjacent membranes. Different versions of the devices were fabricated by changing the membrane thickness from 270 nm to 11.4 μm. To characterize the flatness and coplanarity of the emitter and receiver membranes we used laser scanning confocal microscopy, which allowed us to quantify small height variations with a resolution of 60 nm in the vertical direction. The height profile (the deviation from planarity) along a line traversing the emitter and receiver membranes is shown in Extended Data Fig. 2a–d for membranes with a thickness of 270 nm, 486 nm, 6,712 nm and 11,405 nm. As shown in Extended Data Fig. 2a, b, the flatness deviates by no more than 150 nm over the 80 μm depth of the two thinner devices. For the two thicker devices shown in Extended Data Fig. 2c, d, deviations from flatness can be as large as 400 nm over the 80 μm depth owing to film stresses in the additional SiN layers deposited by PECVD on these devices (see 'Fabrication of the devices'). Compared to the thickness of the membranes, the relative flatness (deviation from planarity/film thickness) of the two thicker devices is better than that of the two thinner devices. Moreover, as discussed in Supplementary Information (see Supplementary Figs. 17, 18), our analysis shows that small deviations from planarity have negligible impact on the RHT. Therefore, the large RHT enhancements in comparison to basic FF-RHT theory are not due to the small imperfections in our devices.

Characterization of the electrical and thermal properties of the devices. Knowledge of the temperature coefficient of resistance of the integrated Pt resistance thermometer (PRT), the thermal frequency response of the device, the beam thermal conductance, and the noise floor of the measurement scheme is crucial for accurate quantification of the radiative exchange between the adjacent membranes. These characterization experiments were performed inside a cryostat (Janis ST-100) evacuated to about 10^{-6} torr using a turbo-molecular pump.

Temperature coefficient of resistance. The resistance (R) of the integrated PRT was measured by passing a small a.c. current of amplitude $I_f = 1\,\mu$A (with frequency $f = 101$ Hz) through the device, resulting in negligible heating. The $1f$ component of the voltage signal across the PRT (V_f) was measured with a lock-in amplifier (bandwidth of about 78 mHz) by using a four-probe scheme, and the resistance of the PRT was calculated as $R = V_f/I_f$. A representative plot of the resistance measured at various temperatures is shown in Extended Data Fig. 3a (the inset shows a schematic of the measurement scheme). We note that a commercial proportional-integral-derivative temperature controller was used to maintain the temperature of the cold finger of the cryostat to within ±10 mK of a fixed point during the experiments at each temperature. The slope of the best-fit line for these data points was used to find the temperature coefficient of resistance, which can be calculated at each temperature as $\alpha(T) = 1/[R(T)] \times \mathrm{d}R/\mathrm{d}T$. At room temperature, the temperature coefficient of resistance was found to be $1.80 \times 10^{-3} \pm 0.03 \times 10^{-3}\ \mathrm{K}^{-1}$ (95% confidence level).

Thermal frequency response. To characterize the frequency response of the devices, their suspended membranes were heated by passing a sinusoidal current of fixed amplitude I_f through the integrated PRT. This sinusoidal current produced Joule heating of the membrane at a frequency of $2f$, which resulted in temperature oscillations of the suspended device at a frequency of $2f$ and an amplitude of ΔT_{2f}. The voltage oscillation across the PRT at frequency $3f$ (V_{3f}) was measured using an SR830 lock-in amplifier (lock-in bandwidth of about 3 mHz). The amplitude of the temperature oscillations was determined from $\Delta T_{2f} = 2V_{3f}/(I_f R\alpha)$. The measured ΔT_{2f} as a function of frequency is shown in Extended Data Fig. 3b for devices of various thicknesses. We note that the a.c. temperature response (ΔT_{2f}) was normalized by the values obtained at low frequencies, which were about 1 K for each device (achieved by suitably adjusting the current I_f). It can be seen that the cut-off frequency depends on the device thickness, and there is negligible attenuation below about 1 Hz for the three thinner devices and about 0.3 Hz for the two thicker devices. At low temperatures, the thermal cut-off frequency was found to shift slightly to higher frequencies.

Beam thermal conductance. To determine the thermal conductance of the beams that support the suspended membrane devices, a large a.c. current (amplitude I_f) was supplied through the integrated PRT at a fixed frequency f, which was selected to be below the thermal cut-off frequency of the device to ensure a full thermal response. The temperature modulation amplitude ΔT_{2f} was obtained from the $3f$ component of the voltage across the PRT, as described previously. The modulated heat input was calculated as $Q_{2f} = I_f^2 R_{\mathrm{tot}}/2$, where R_{tot} is the total electrical resistance of the integrated PRT and the current-carrying leads on the beams. The current-carrying leads were found to have a small resistance, of about 100 Ω, compared to the resistance of the PRT (about 10 kΩ); hence, heating in the beams has negligible contributions to the heat input into the membrane[23]. Extended Data Fig. 3c shows the amplitude of the temperature oscillations ΔT_{2f} as a function of the heating power Q_{2f}. The inverse of the slope of the best-fit line for these data points provides the thermal conductance of the beams. As expected, the thermal conductance depends on the thickness of the beams and at room temperature ranges from about 80 nW K^{-1} for the thinnest device (270 nm thick) to about 1 μW K^{-1} for the thickest device (11.4 μm thick). The thermal conductance decreases at lower temperatures owing to the temperature-dependent thermal conductivity of silicon nitride. We note that for the thinner devices there are substantial contributions to the thermal conductance (about 75%) from the metal lines integrated into it. We also note that for the two thicker devices we have two different kinds of SiN (deposited by LPCVD and by PECVD), whereas for the three thinner devices we only used LPCVD-deposited SiN. Past work[33,34] suggested that PECVD-deposited SiN has a lower thermal conductivity than LPCVD-deposited SiN, which partly explains why the thermal cut-off frequency is lower for the two thicker devices.

Measurement of $G_{rad,eff}$. The effective radiative conductance between adjacent membranes was measured by first applying a large a.c. current (amplitude I_f) through the integrated PRT of the emitter membrane. The frequency of the a.c. current was selected to ensure a full thermal response of the device (that is, f was chosen to be well below the cut-off frequency for each of the devices). The temperature modulation amplitude of the emitter membrane, $\Delta T_{\mathrm{emit},2f}$, was found from the $3f$ component of the voltage across the PRT, as described previously. The resulting radiative heat flow ($Q_{\mathrm{rad},2f}$) to the receiver membrane produced a modulated temperature rise ($\Delta T_{\mathrm{rec},2f}$) in the receiver membrane. A d.c. current (amplitude I_{dc}) was supplied through the PRT to the receiver membrane, and the $2f$ component of the voltage across the PRT (V_{2f}) was used to calculated the temperature modulation amplitude as $\Delta T_{\mathrm{rec},2f} = V_{2f}/(I_{\mathrm{dc}}R\alpha)$. A lock-in amplifier, which had a bandwidth of 0.78 mHz to reduce noise, was used to detect V_{2f}. More specifically, $\Delta T_{\mathrm{emit},2f}$ and $\Delta T_{\mathrm{rec},2f}$ (Fig. 2a, Extended Data Fig. 3d) were obtained by measuring the lock-in output every 2 s (a total of 1,500 measurements) and computing the mean of the data. Error bars for $\Delta T_{\mathrm{emit},2f}$ and $\Delta T_{\mathrm{rec},2f}$ were calculated from the standard deviation of the same data. The measured $G_{\mathrm{rad,eff}}$ values plotted in Figs. 2, 3 were determined by recording $\Delta T_{\mathrm{rec},2f}$ for various values of $\Delta T_{\mathrm{emit},2f}$ and calculating the slope of the linear regression line for these data (a representative plot is shown in Fig. 2a). Finally, the effective radiative conductance between the membranes is obtained from $G_{\mathrm{rad,eff}} = G_{\mathrm{beams}}\Delta T_{\mathrm{rec},2f}/(\Delta T_{\mathrm{emit},2f} - \Delta T_{\mathrm{rec},2f})$, whereas the error bars for $G_{\mathrm{rad,eff}}$ were estimated from the uncertainty in the slope of the linear model described above.

Noise floor of the measurement scheme. To estimate the noise floor of the measurement scheme, the emitter membrane was Joule-heated with progressively lower power, and $\Delta T_{\mathrm{emit},2f}$ and $\Delta T_{\mathrm{rec},2f}$ were simultaneously measured at each power using the procedure described previously. Representative data are shown in Extended Data Fig. 3d for a 2-μm-thick device at 100 K for $f = 0.5$ Hz, a bandwidth of 0.78 mHz and $I_{\mathrm{dc}} = 10\,\mu$A (lower temperature represents a more challenging scenario because the radiative conductance is smaller). It can be seen that $\Delta T_{\mathrm{rec},2f}$ decreases linearly for decreasing $\Delta T_{\mathrm{emit},2f}$. When $\Delta T_{\mathrm{emit},2f}$ is less than about 1 K, the signal measured by the receiver becomes noise limited: that is, for a progressively smaller temperature increase in the emitter membrane, the measured temperature rise of the receiver membrane never falls below 60 μK, which represents the noise-equivalent temperature (NET). Similar experiments performed at other temperatures reveal that the NET falls into the range 25–70 μK depending on the choice of the modulation frequency and the magnitude of the d.c. sensing current I_{dc}. By knowing the NET and the thermal conductance of the beam of each device, we could estimate the heat flow resolution as $Q_{\mathrm{res}} = G_{\mathrm{beams}} \times \mathrm{NET}$ at any given temperature. We used this estimation of the heat flow resolution to ensure that none of the measurements of the effective radiative conductance were limited by noise. Typically, this required us to set $\Delta T_{\mathrm{emit},2f}$ to be about 5–10 K.

Control experiments. To rule out that transport mechanisms other than far-field radiation or measurement artefacts contribute to the reported conductances, we conducted several control experiments. Such parasitic coupling mechanisms between the membranes may include conduction through residual air molecules inside the cryostat chamber, conduction through the Si handle substrate in which the membranes are embedded, radiative exchange between the membranes that is mediated by reflections from various surfaces within the vacuum chamber, and capacitive or inductive coupling of the a.c. current passing through the Pt heater on the emitter membrane to the Pt lines on the receiver membrane.

Conduction through residual air molecules. A simple analysis based on the kinetic theory of gases suggests that at this pressure, energy exchange between the membranes due to conduction through residual air molecules is negligible compared to the measured radiative coupling. To experimentally verify this hypothesis, we measured $G_{\mathrm{rad,eff}}$ at two different pressures (10^{-6} torr and 10^{-3} torr). Because conduction through air molecules is proportional to the pressure (at sufficiently low pressures)[35], increasing the pressure from 10^{-6} torr to 10^{-3} torr should result in a 1,000-fold increase in this exchange mechanism. Thus, if parasitic conduction through air molecules had sizeable contributions to $G_{\mathrm{rad,eff}}$ at 10^{-6} torr, then at 10^{-3} torr we should measure a large increase in the effective conductance between the emitter and receiver. Instead, we found no measurable increase. Results from measurements on a device featuring 2-μm-thick membranes (Extended Data Fig. 4a), demonstrate that the measured conductance is independent of the vacuum level for pressures between 10^{-3} torr and 10^{-6} torr. This independence from pressure was confirmed for cryostat temperatures from 80 K to 300 K, suggesting that conduction through residual air molecules at 10^{-6} torr was indeed negligible in all our experiments and can be ignored.

Conduction through the handle substrate. All device structures were fabricated using thin SiN films deposited on silicon handle wafers (500 μm thick). Therefore, one could envision a scenario in which heat from the emitter membrane conducts out through the support beams and the silicon handle wafer and into the receiver membrane through its support beams, producing a measurable signal that could be misinterpreted as radiative gap conductance (see inset of Extended Data Fig. 4b). To test this hypothesis, a set of membranes was fabricated with short support beams (150 μm long instead of 400 μm long). We expected that an increase in the measured gap conductance between these short-beam devices compared to that between devices with longer beams would signal a non-negligible energy exchange via conduction through the substrate. However, as shown in Extended Data Fig. 4b, the gap conductance measured between these short-beam devices is essentially equal to that measured for the devices with 400-μm-long beams. This control experiment verifies that conduction through the Si handle substrate has a negligible contribution to the measured gap conductance and cannot account for the measured orders-of-magnitude enhancements in the gap conductance.

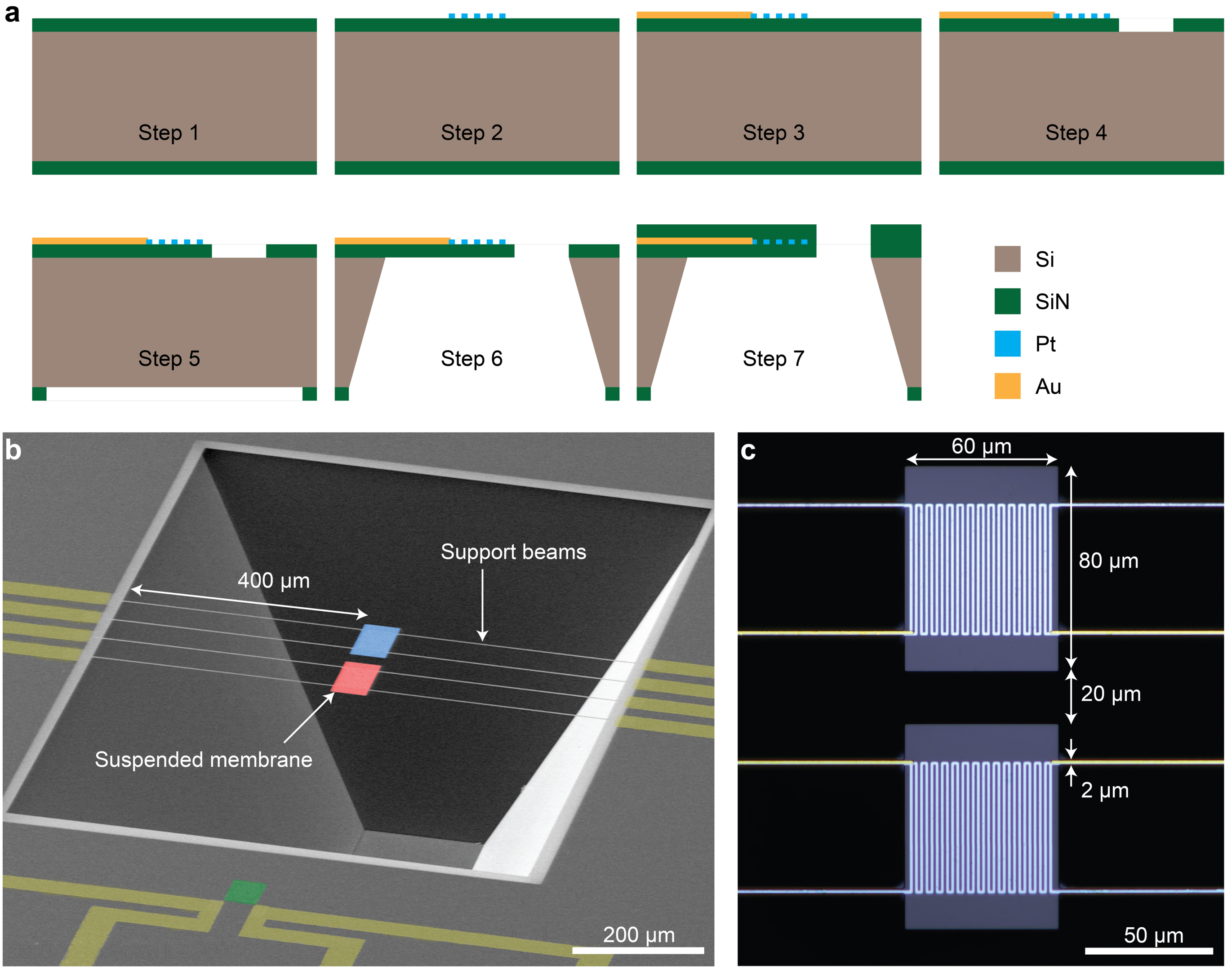
Demonstration of radiative coupling. To further confirm that heat transport from the emitter to the receiver is completely dominated by direct RHT from the emitter to the receiver, we prepared disconnected emitter and receiver devices by cleaving the original (paired) devices (membrane thickness of 2 μm) in the middle. Next, we placed a separated emitter at a distance of about 1 mm from a single receiver (see Extended Data Fig. 4c). Confocal microscopy was performed to ensure that the devices were coplanar with each other. Next, a thin shim of reflective aluminium foil (thickness 24 μm) was placed in the gap between adjacent membranes, blocking direct radiative exchange between them, but allowing possible radiative exchange via reflections from the top side (see inset of Extended Data Fig. 4d) and conduction through the underlying support structure. As shown in Extended Data Fig. 4d, with the aluminium shim in place, the measured coupling signal between the membranes is completely blocked (red). For membranes separated by a gap of the same size but without the aluminium foil shim in place (black), there is a linear increase in the temperature of the receiver membrane as the temperature of the emitter membrane is raised. This control test conclusively illustrates that heat transfer between membranes is completely dominated by direct RHT from the emitter to the receiver device. We also note that in all our experiments, the devices were mounted over Metal Velvet foil with a hemispherical reflectance of less than 4% for wavelengths of less than 14 μm. Further, the inner surface of the radiation shield of the cryostat, which the membranes 'see' on the top side, was also coated with Metal Velvet foil. As verified by COMSOL simulations (and the control experiment described earlier), such a mounting scheme effectively eliminates radiative coupling via reflections from the cold-finger surface and the radiation shield.

Proof of negligible coupling via capacitive and inductive effects or via reflection of radiation. For each measurement of the effective radiative conductance between the emitter and receiver devices, the temperature of the emitter membrane was modulated by providing an a.c. current to the integrated Pt heater. Any capacitive coupling[36] between the Pt lines on the emitter and the Pt lines on the receiver due to this a.c. current and its harmonics could potentially produce spurious signals that might be misinterpreted as radiative conductance signals. Control experiments were performed to quantify the magnitude of this capacitive coupling. In these experiments we took advantage of the fact that the suspended devices used in the experiments were fabricated on a 10 mm × 10 mm silicon handle chip and each chip had four emitter/receiver pairs separated by 20-μm gaps, with each pair suspended in a separate through-hole in the handle chip (see Extended Data Fig. 4e, which shows two emitter/receiver pairs in separate through-holes). To quantify the contributions from capacitive coupling, we used an emitter from one pair of devices and a receiver from a second pair of devices (see Extended Data Fig. 4e). In this configuration, direct radiative exchange between the emitter and receiver is blocked, but capacitive coupling between these non-adjacent devices is still possible because the electrical leads and contact pads on the handle substrate are in close proximity. As shown in Extended Data Fig. 4f, the coupling signal between non-adjacent devices in separate through-holes (red circles) is extremely small and is noise-limited. For comparison, the coupling signal between adjacent devices separated by a 20-μm gap (black squares) is shown to be readily measurable. This control test verifies that capacitive coupling between adjacent membranes is indeed negligible. Further, since there is no measurable coupling between the emitter from the first pair and the receiver from the second pair, one can also conclude that any radiative coupling via reflections within the cryostat is also negligible.

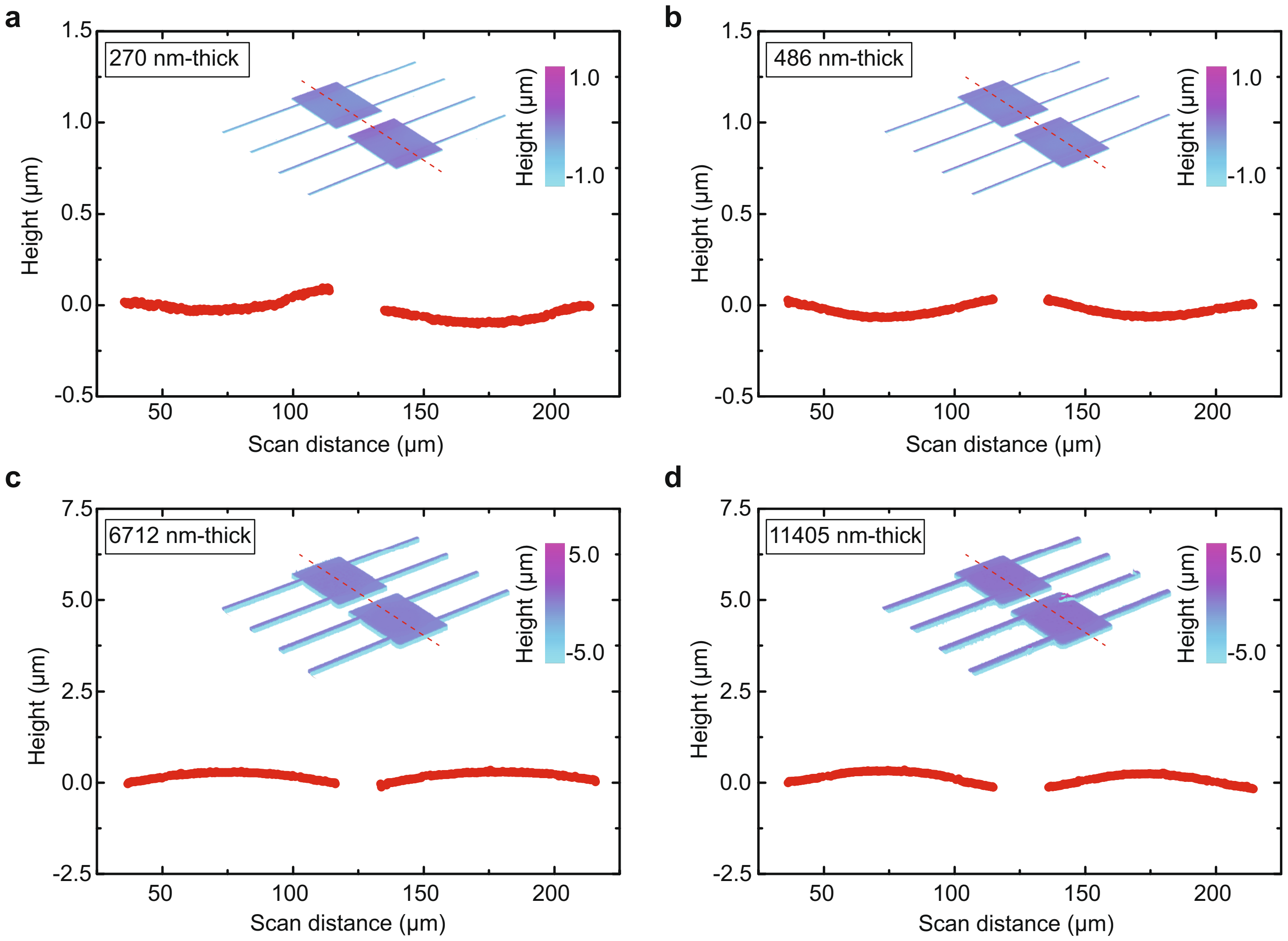
Computational methods. Calculations of RHT in the framework of FF-RHT theory were performed using COMSOL Multiphysics. Calculations of RHT within the FED framework were performed using a BEM[27,30]. An implementation of this method is available via the open-source code SCUFF-EM (https://homerreid.github.io/scuff-em-documentation). More details about the computational methods, including mesh sizes, convergence criterion and analysis of different materials and structures, are provided in Supplementary Information.

Data availability. The data that support the findings of this study are available from the corresponding authors on reasonable request.

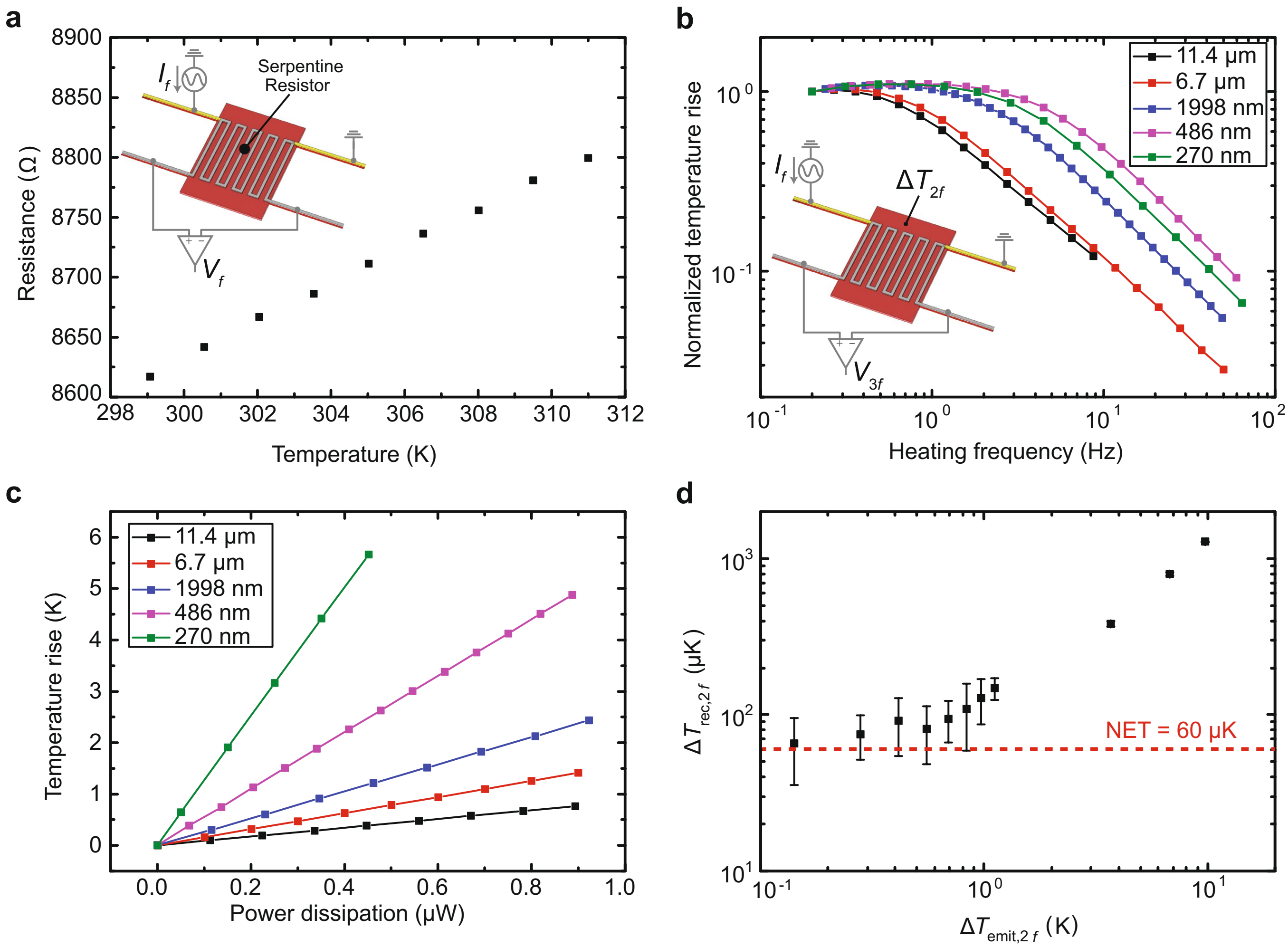
33. Sultan, R., Avery, A., Stiehl, G. & Zink, B. Thermal conductivity of micromachined low-stress silicon-nitride beams from 77 to 325 K. *J. Appl. Phys.* **105**, 043501 (2009).
34. Lee, S.-M. & Cahill, D. G. Heat transport in thin dielectric films. *J. Appl. Phys.* **81**, 2590–2595 (1997).
35. Kennard, E. H. *Kinetic Theory of Gases, with an Introduction to Statistical Mechanics* (McGraw-Hill, New York, 1938).
36. Horowitz, P. & Hill, W. *The Art of Electronics* (Cambridge Univ. Press, Cambridge, 1989).


a
Step 1
Step 2
Step 3
Step 4
Step 5
Step 6
Step 7
Si
SiN
Pt
Au
b
400 µm
Support beams
Suspended membrane
200 µm
c
60 µm
80 µm
20 µm
2 µm
50 µm

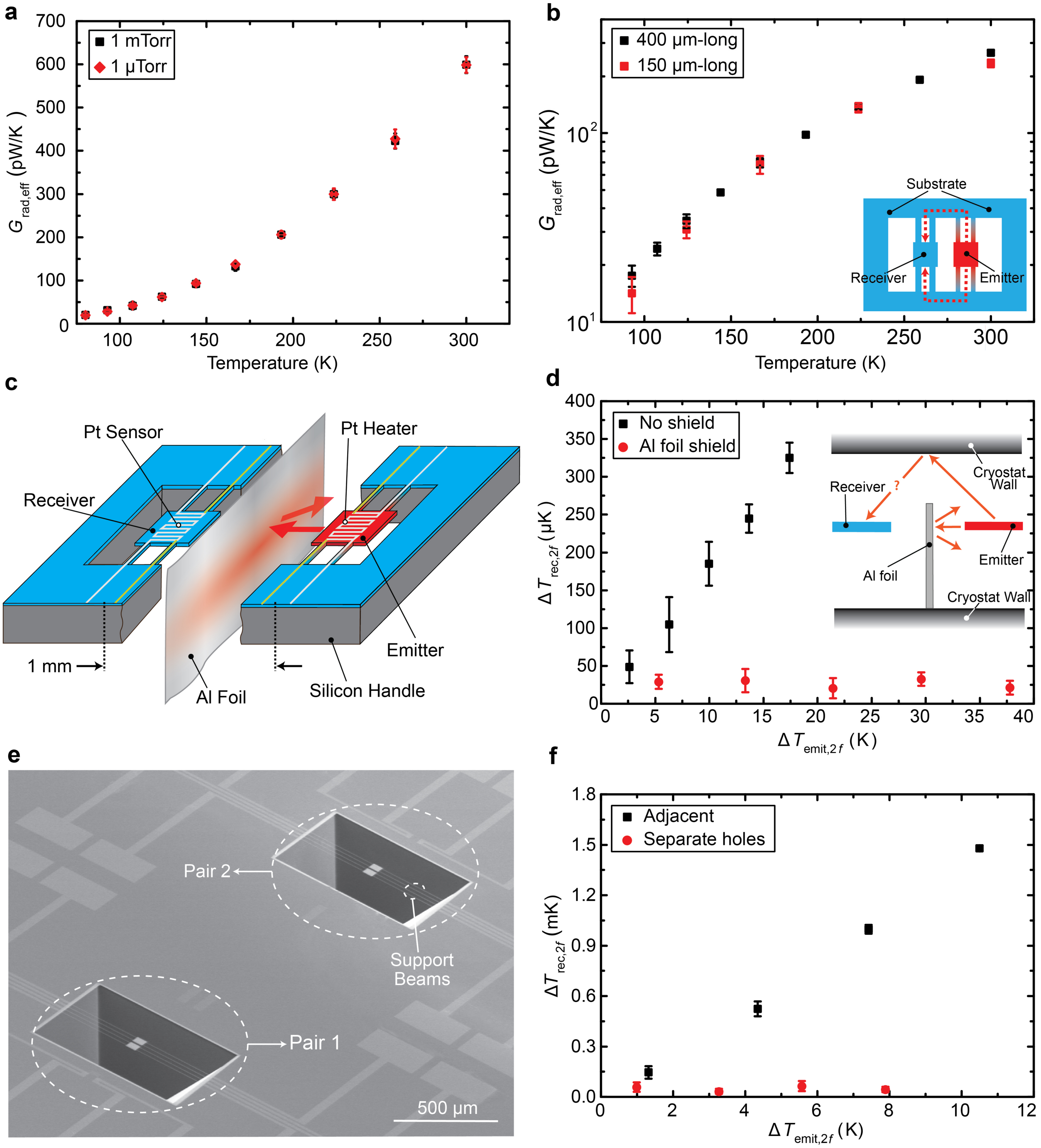
Extended Data Fig. 1 | Fabrication of the devices. a, Step 1: deposition of the silicon nitride (SiN) layer on a silicon handle wafer. Step 2: patterning of the Pt serpentine resistance thermometer. Step 3: patterning of gold leads. Step 4: front-side etching of suspended device contour. Step 5: etching of a window in SiN on the back side of the device. Step 6: potassium hydroxide (KOH) etching of silicon handle to release the suspended devices. Step 7: optional PECVD of SiN onto suspended 1,998-nm-thick devices to create even thicker membranes (6,712 and 11,405 nm thick). **b**, **c**, Scanning electron microscope (**b**) and optical microscope (**c**) images showing the geometry of the fabricated devices and relevant dimensions. Au wires (yellow) and suspended membranes (red and blue) were pseudo-coloured.



Extended Data Fig. 2 | Characterization of the flatness and coplanarity of adjacent membranes. a–d, Data from laser scanning confocal microscope scans across adjacent membranes showing the flatness of 270-nm-thick (**a**), 486-nm-thick (**b**), 6,712-nm-thick (**c**) and 11,405-nm-thick (**d**) membranes. Line profiles were taken over the red dashed line indicated in the surface plots in the insets.

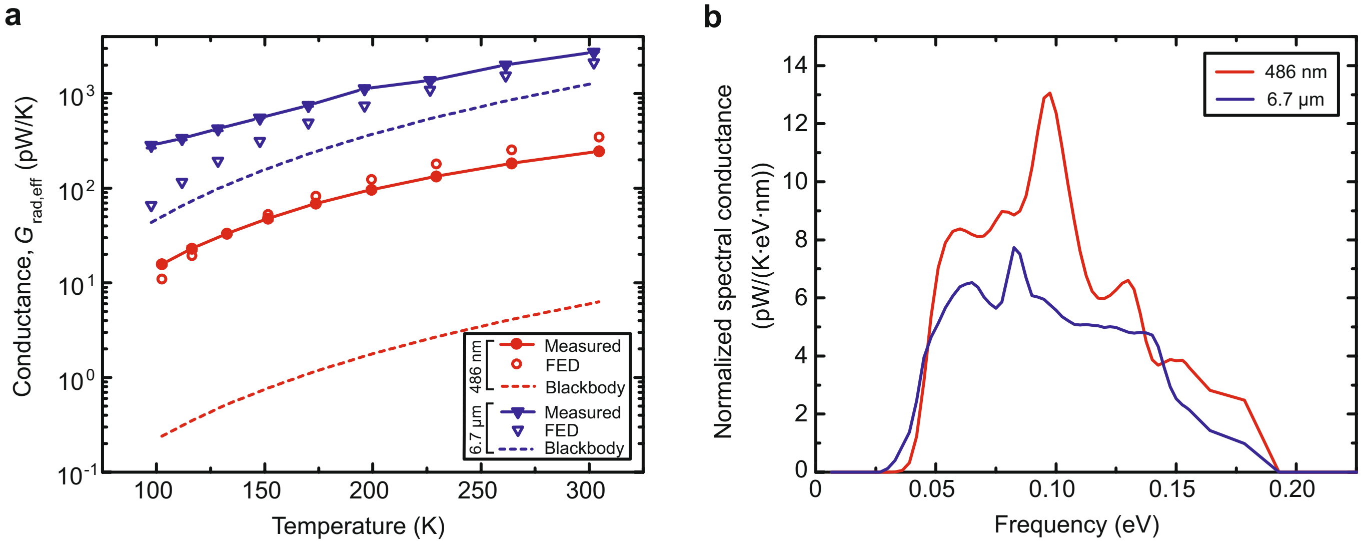


Extended Data Fig. 3 | Characterization of the electrical and thermal properties of the devices. a, Temperature dependence of the resistance of the integrated PRT near room temperature. The inset shows a schematic of the measurement scheme used for resistance characterization. **b**, Thermal frequency response of membrane devices with various thicknesses measured at $T = 300$ K. The inset depicts a schematic of the measurement scheme employed for these tests. **c**, Temperature rise of the membrane as a function of the Joule heating in the PRT, measured at $T = 300$ K for devices of various thicknesses. **d**, Temperature rise of the receiver membrane as a function of the temperature increase of the emitter membrane for a 2-μm-thick device at 100 K. The noise-equivalent temperature is about 60 μK for a temperature modulation frequency of 0.5 Hz and a measurement bandwidth of 0.78 mHz, and $I_{\mathrm{dc}} = 10$ μA.



Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | Quantification of parasitic energy transfer mechanisms. a, Measured $G_{\mathrm{rad,eff}}$ between 2-μm-thick SiN membranes at various temperatures and pressures. Measurements performed at 10^{-3} torr (black squares) and 10^{-6} torr (red diamonds) give virtually identical results, strongly supporting the expectation that conduction via remaining gas molecules plays no role. **b**, Measured $G_{\mathrm{rad,eff}}$ between 270-nm-thick SiN membranes at various temperatures. The inset illustrates how thermal energy between the devices can potentially couple via conduction through the substrate. The plotted data correspond to two device geometries: devices with 400-μm-long support beams (black squares) and devices with 150-μm-long support beams (red squares). Because devices with short and long beams have identical $G_{\mathrm{rad,eff}}$, it is clear that coupling via the suspending beams is negligibly small. **c**, Schematic describing the control experiment, where an emitter and a receiver are separated by a gap of about 1 mm and an Al foil is placed between them, blocking direct RHT (not drawn to scale and proportion). **d**, Measured coupling signal between 2-μm-thick emitter and receiver membranes across a 1-mm-wide gap ($T = 300$ K). Signals without (black squares) and with (red circles) an Al foil in the gap are shown. The inset (cross-sectional view of the devices) illustrates how the Al foil shield blocks direct RHT between the membranes but potentially allows RHT via specular reflections. Because the signal obtained in the presence of the Al blocker represents the noise floor of our measurement, we conclude that there is negligible heat transfer (via RHT or otherwise) in the presence of the Al foil. This control experiment provides unequivocal evidence that energy transfer between the membranes is mediated exclusively by direct radiation. **e**, Scanning electron microscope image of two emitter and receiver pairs suspended in separate through-holes on the same handle substrate. **f**, Measured coupling signal between 486-nm-thick devices ($T = 100$ K). Signals from adjacent emitter and receiver devices (that is, in the same through-hole; see pair 1 in **e**) with a 20-μm-gap (black squares) are compared with the signals measured when the emitter and receiver devices are suspended in separate through-holes on the same silicon handle substrate (red circles).



Extended Data Fig. 5 | Radiative conductance for devices with 486-nm-thick and 6,712-nm-thick membranes. a, $G_{\mathrm{rad,eff}}$ as a function of temperature. Experimental measurements (solid symbols) are compared with values calculated using FED (open symbols) and COMSOL-modelled blackbody values (dashed lines). **b**, BEM calculation of the normalized spectral radiative conductance at 300 K. The spectral conductance values are normalized by the thickness of each device.

LETTER

https://doi.org/10.1038/s41586-018-0490-7

Tunable long-distance spin transport in a crystalline antiferromagnetic iron oxide

R. Lebrun[1,6]*, A. Ross[1,2,6], S. A. Bender[3], A. Qaiumzadeh[4], L. Baldrati[1], J. Cramer[1,2], A. Brataas[4], R. A. Duine[3,4,5] & M. Kläui[1,2,4]*

Spintronics relies on the transport of spins, the intrinsic angular momentum of electrons, as an alternative to the transport of electron charge as in conventional electronics. The long-term goal of spintronics research is to develop spin-based, low-dissipation computing-technology devices. Recently, long-distance transport of a spin current was demonstrated across ferromagnetic insulators[1]. However, antiferromagnetically ordered materials, the most common class of magnetic materials, have several crucial advantages over ferromagnetic systems for spintronics applications[2]: antiferromagnets have no net magnetic moment, making them stable and impervious to external fields, and can be operated at terahertz-scale frequencies[3]. Although the properties of antiferromagnets are desirable for spin transport[4–7], indirect observations of such transport indicate that spin transmission through antiferromagnets is limited to only a few nanometres[8–10]. Here we demonstrate long-distance propagation of spin currents through a single crystal of the antiferromagnetic insulator haematite (α-Fe_2O_3)[11], the most common antiferromagnetic iron oxide, by exploiting the spin Hall effect for spin injection. We control the flow of spin current across a haematite–platinum interface—at which spins accumulate, generating the spin current—by tuning the antiferromagnetic resonance frequency using an external magnetic field[12]. We find that this simple antiferromagnetic insulator conveys spin information parallel to the antiferromagnetic Néel order over distances of more than tens of micrometres. This mechanism transports spins as efficiently as the most promising complex ferromagnets[1]. Our results pave the way to electrically tunable, ultrafast, low-power, antiferromagnetic-insulator-based spin-logic devices[6,13] that operate without magnetic fields at room temperature.

Spin-wave excitations of magnetic moments, called magnons, can transport spin angular momentum in ferromagnetic and antiferromagnetic insulators. In easy-axis antiferromagnets, the two degenerate magnon modes have left or right circular polarization, carrying finite, but opposite, angular momenta[14]. At thermal equilibrium, there is no net transport because these degenerate, equal-frequency magnons cannot be populated separately. However, in bilayer devices consisting of an antiferromagnetic insulator and a heavy metal, a spin accumulation at the interface between the two layers can generate an imbalance of the magnon population, enabling spin transport[12,15]. Such a spin bias, along either the Néel vector $\boldsymbol{n}$ or the field-induced magnetization $\boldsymbol{m}$, can efficiently excite or annihilate the magnon modes depending on the sign of the bias. This potentially enables the antiferromagnetic spin conductance to be probed, because the magnon modes independently transport spin angular momentum. In parallel, thermal excitations by Joule heating can propagate magnons through the spin-Seebeck conductance, in the presence of a field-induced net magnetization $\boldsymbol{m}$[16,17]. Low magnetic damping has been reported for antiferromagnetic insulators[3,18], so these materials lend themselves naturally to efficient long-distance spin-transport even though such transport has not been observed previously[8–10].

Signatures of long-distance spin-superfluid transport, rather than diffusive transport, through an antiferromagnet were reported recently[19], but were based on only thermal transport signals, without any detection of the expected accompanying spin-injected transport signal. As discussed recently[20,21], other effects such as spatially extended thermal gradients can explain non-local thermal signals. By achieving long-distance spin transport by spin injection, we unambiguously demonstrate that such transport is consistent with diffusive rather than superfluid transport, being accompanied by a thermal signal that decays over even longer distances than does the transport signal (as discussed below and in Methods). Our results clearly rule out spin-superfluidity, but demonstrate that long-distance, diffusive spin transport through an antiferromagnet is possible even at elevated temperatures (above 200 K).

To investigate the spin-transport mechanisms in antiferromagnetic insulators, we use a non-local geometry[1] of platinum wires on a single crystal of α-Fe_2O_3 (Fig. 1a). A charge current $\boldsymbol{I}$ passes through a platinum wire, inducing two effects. First, the spin Hall effect (SHE) produces a transverse spin current, which leads to a spin accumulation $\boldsymbol{\mu}$ at the Pt–α-Fe_2O_3 interface[22]. This accumulation may couple to the antiferromagnetic order, generating a spin current that carries a net angular momentum (Fig. 1a). Second, Joule heating of the platinum wire induces a lateral temperature gradient ΔT, which generates a spin-Seebeck-induced thermal spin current. Ultimately, the non-local voltage $V = V_{\mathrm{el}} + V_{\mathrm{th}}$, where V_{el} originates from the injected spin current and V_{th} from thermal effects, is a combination of both spin-current effects, which are detected at a non-local platinum wire via the inverse SHE[22]. The even component of V, which is quadratic in current and related to the spin-Seebeck conductance S, can be determined from the voltage measured for the two current polarities (I^+ and I^-) as $V_{\mathrm{th}} = [V(I^+) + V(I^-)]/2$, and the odd component, which is linear in current and describes the spin conductance G, from $V_{\mathrm{el}} = [V(I^+) - V(I^-)]/2$, to remove any thermal contributions[20]. The non-local signals V_{th} and V_{el} arise from spin currents that carry angular momentum along the antiferromagnetic Néel order $\boldsymbol{n} = (\boldsymbol{m}_1 - \boldsymbol{m}_2)/2$ and magnetic-field-induced moment $\boldsymbol{m} = (\boldsymbol{m}_1 + \boldsymbol{m}_2)/2$, where $\boldsymbol{m}_1$ and $\boldsymbol{m}_2$ are the two sublattice magnetizations. We therefore write the non-local resistances $R_{\mathrm{el}} = V_{\mathrm{el}}/I$ (in units of Ω) and $R_{\mathrm{th}} = V_{\mathrm{th}}/I^2$ (in units of V A^{-2}) using a two-channel, phenomenological model (Methods):

$$\begin{aligned} R_{\mathrm{el}} &= G_{\boldsymbol{n}}(\boldsymbol{n}\cdot\boldsymbol{e}_\perp)^2 + G_{\boldsymbol{m}}(\boldsymbol{m}\cdot\boldsymbol{e}_\perp)^2 \\ R_{\mathrm{th}} &= S_{\boldsymbol{n}}(\boldsymbol{n}\cdot\boldsymbol{e}_\perp) + S_{\boldsymbol{m}}(\boldsymbol{m}\cdot\boldsymbol{e}_\perp) \end{aligned} \tag{1}$$

where $\boldsymbol{e}_\perp$ is a unit vector normal to the platinum wires, that is, parallel to the current-induced $\boldsymbol{\mu}$. All four coefficients ($G_{\boldsymbol{n}}$, $G_{\boldsymbol{m}}$, $S_{\boldsymbol{n}}$, $S_{\boldsymbol{m}}$) depend on the directions of $\boldsymbol{n}$ and $\boldsymbol{m}$. Finally, R_{el}, and therefore $G_{\boldsymbol{n}}$ and $G_{\boldsymbol{m}}$, are determined from direct measurements of the spin conductance of the antiferromagnet, which decreases with the distance from the injector to the detector[1].

Our study requires full control of $\boldsymbol{n}$, which cannot be achieved by aligning the magnetization with small external fields such as in

---

[1]Institute for Physics, Johannes Gutenberg-University Mainz, Mainz, Germany. [2]Graduate School of Excellence Materials Science in Mainz, Mainz, Germany. [3]Utrecht University, Utrecht, The Netherlands. [4]Center for Quantum Spintronics, Department of Physics, Norwegian University of Science and Technology, Trondheim, Norway. [5]Department of Applied Physics, Eindhoven University of Technology, Eindhoven, The Netherlands. [6]These authors contributed equally: R. Lebrun, A. Ross. *e-mail: rolebrun@uni-mainz.de; klaeui@uni-mainz.de

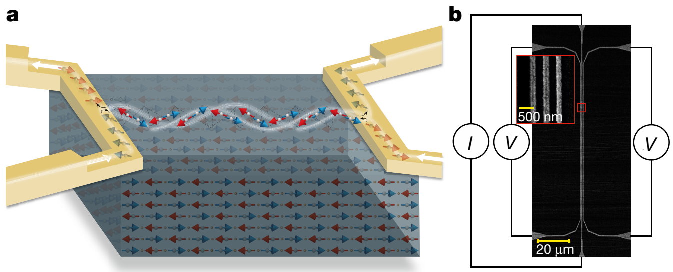


Fig. 1 | Spin transport in an insulating antiferromagnet. a, Schematic of the experiment, in which two platinum wires (yellow) are placed on an insulating easy-axis antiferromagnet (α-Fe_2O_3, grey; red and blue arrows illustrate oppositely oriented spins). The SHE (white arrows) in the left wire generates a spin accumulation at the Pt–α-Fe_2O_3 interface, breaking the antiferromagnetic symmetry. By transferring angular momentum to the antiferromagnet (black arrows), this symmetry breaking excites magnons, which diffuse to the right wire (artistically shown as white shaded lines), where the spin current is absorbed (black arrows on the right) and detected via the inverse SHE (white arrows on the right). The dashed lines show the precession of the sublattice spins. **b**, Scanning electron microscopy (SEM) image of a typical device, with wire spacings of 200 nm and 250 nm. The platinum injector and detector wires are connected to chromium/gold contacts; current I and voltage V connections are indicated.

ferromagnets. However, in low-anisotropy antiferromagnets such as α-Fe_2O_3, the direction of $\boldsymbol{n}$ can still be controlled using a field. Above the spin-flop field $\boldsymbol{H}_c$ (of about 6 T at 200 K)[18], $\boldsymbol{n}$ reorients perpendicular to the applied field $\boldsymbol{H}$. We control the antiferromagnetic order by sweeping an in-plane magnetic field along different directions (Methods) to explore the spin-transport signal in devices with different injector–detector distances (Fig. 1b, Methods).

We first consider a device geometry with platinum wires oriented along $\boldsymbol{x}$, the in-plane axis onto which the easy axis is projected (Methods). Initially, $\boldsymbol{n}$ is approximately perpendicular to $\boldsymbol{\mu} = \mu\boldsymbol{y}$. As we sweep $\boldsymbol{H}$ along $\boldsymbol{x}$, $\boldsymbol{n}$ rotates smoothly and becomes perpendicular to $\boldsymbol{H}$ (along $\boldsymbol{y}$) when the field reaches $\boldsymbol{H}_c$. As illustrated in Fig. 2a, we find that the spin-conductance signal R_{el} exhibits a maximum at $\boldsymbol{H}_c$ (around 6 T) and remains non-zero at larger fields, while R_{th} is zero. For $H > H_c$, R_{el} is non-zero owing to the reorientation of $\boldsymbol{n}$ along $\boldsymbol{\mu} = \mu\boldsymbol{y}$ (see sketch in Fig. 2a). Therefore, we identify the Néel spin conductance $G_{\boldsymbol{n}}$ as the primary mechanism for carrying angular momentum; the spin accumulation along $\boldsymbol{n}$ excites the antiferromagnetic magnon mode with polarization along $\boldsymbol{\mu}$, while annihilating the other mode. This mechanism also explains the sharp peak in R_{el} at $\boldsymbol{H}_c$, when the field compensates the anisotropy energy and the magnon gap of one mode greatly decreases[18]. A full gap closure theoretically leads to a divergence of $G_{\boldsymbol{n}}$. On the basis of this explanation, we modelled the experiment as shown in Fig. 2. We find that for our geometry the magnon gap of one mode is reduced by a factor of ten, illustrating that $G_{\boldsymbol{n}}$ depends on both the magnon gap and the direction of $\boldsymbol{n}$, and can thus be tailored. In addition, we find that the Néel spin-Seebeck conductance $S_{\boldsymbol{n}}$ is negligible, resulting in the absence of R_{th} below and above $\boldsymbol{H}_c$.

We then study a second field direction, with $\boldsymbol{H}$ applied along $\boldsymbol{y}$ (Fig. 2b). In this case there is a field-induced magnetization $\boldsymbol{m}$ parallel to $\boldsymbol{\mu} = \mu\boldsymbol{y}$ (Fig. 2b). The absence of R_{el} is consistent with the prediction that the magnetic spin conductance $G_{\boldsymbol{m}}$ should be reduced by a factor of $T\chi \ll 1$ compared to $G_{\boldsymbol{n}}$ (with T the temperature and χ the susceptibility)[12]. However, the spin-Seebeck signal R_{th} now contributes, increasing linearly with $\boldsymbol{H}$. S is non-zero only in the presence of $\boldsymbol{m}$ along $\boldsymbol{\mu}$, indicating that $S_m \gg S_n$. In antiferromagnets, $\boldsymbol{H}$ does not break the symmetry of $\boldsymbol{n}$ and the spin accumulation is only along $\boldsymbol{m}$; hence, $S_{\boldsymbol{n}}$ vanishes in the absence of sublattice-symmetry breaking.

We initially conclude that a spin current propagates through an antiferromagnetic insulator, mediated by $G_{\boldsymbol{n}}$ and $S_{\boldsymbol{m}}$ (full angular scans with theoretical fits are provided in Methods). Contrary to ferromagnets[1,23], the different spin conductances thus rely on two different magnetic order parameters, $\boldsymbol{n}$ and $\boldsymbol{m}$; therefore, spin-injected and thermal spin currents could be controlled separately only in antiferromagnetic spin-logic devices[13,24].

The application of strong fields is cumbersome, so we consider platinum wires along $\boldsymbol{y}$, perpendicular to the easy-axis at zero applied field (Fig. 3a, b). We first analyse this geometry for $\boldsymbol{H}$ applied

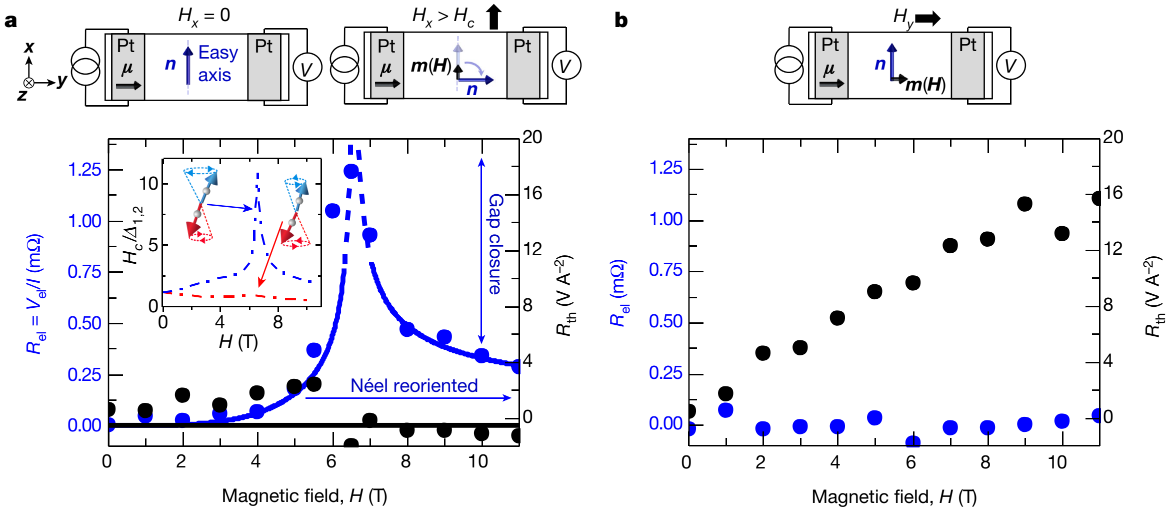


Fig. 2 | Spin transport at 200 K for platinum wires along *x*. a, For a magnetic field $\boldsymbol{H}$ applied parallel to the wires (along $\boldsymbol{x}$), the non-local signal R_{el} (blue circles, left axis) is zero at small fields, exhibits a sharp peak at the spin-flop field $\boldsymbol{H}_c$ and is non-zero above $\boldsymbol{H}_c$. R_{el} is modelled theoretically by the lowest magnon gap Δ and the projection of $\boldsymbol{\mu}$ on $\boldsymbol{n}$ ($\boldsymbol{\mu} \cdot \boldsymbol{n}$) (blue line; the dashed part indicates theoretical divergence at $\boldsymbol{H}_c$; see Methods for more details). The Néel vector $\boldsymbol{n}$ is reoriented perpendicular to $\boldsymbol{H}$ above $\boldsymbol{H}_c$ (in the region indicated by the arrow). At $\boldsymbol{H}_c$, the magnon gap of one mode closes, leading to a large increase in the signal that is supressed as the gap reopens at higher fields (the increase from the gap closing is indicated by an arrow). The inset illustrates how the theoretical inverse gap $1/\Delta$ of one mode ($1/\Delta_1$) is reduced at $\boldsymbol{H}_c$ (red), while the other is enhanced ($1/\Delta_2$; blue). The two modes are represented by the two spin configurations, with cones of precession indicated by the dotted lines. The dynamical opening angle of the two sublattices is exaggerated. The thermal signal R_{th} (black circles, right axis) remains low; the solid black line is from theoretical fitting of both R_{el} and R_{th} (Methods). **b**, With the magnetic field perpendicular to the wires (along $\boldsymbol{y}$), the field-induced magnetization $\boldsymbol{m}$ is parallel to $\boldsymbol{\mu}$ and perpendicular to $\boldsymbol{n}$. R_{el} remains zero and R_{th} increases with the field. The schematics of the experiments (above the plots) demonstrate the relative orientations of $\boldsymbol{\mu}$, $\boldsymbol{n}$ and $\boldsymbol{m}$.

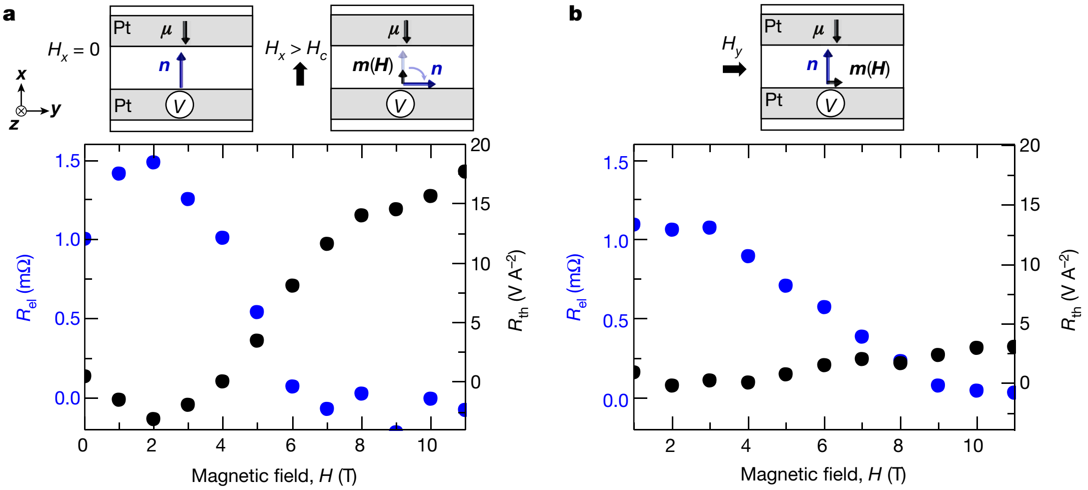


Fig. 3 | Spin transport at 200 K for platinum wires along *y*. **a**, For ***H*** applied (along ***x***) perpendicular to the wires, R_{el} (blue, left axis) is non-zero at low fields, sharply decreasing at $\boldsymbol{H}_c$, when ***n*** aligns perpendicular to the spin accumulation $\boldsymbol{\mu} = \mu\boldsymbol{x}$. R_{th} (black, right axis) is initially zero, enhancing at $\boldsymbol{H}_c$ when a field-induced magnetization ***m*** emerges along $\mu\boldsymbol{x}$. **b**, For ***H*** applied (along ***y***) parallel to the wires, R_{el} decreases with applied magnetic field as ***n*** reorients out of the *x*–*y* plane at large fields. R_{th} remains zero because ***m*** is always perpendicular to $\mu\boldsymbol{x}$.

along ***x*** (Fig. 3a). The striking feature here is the presence of a strong R_{el} signal at zero applied field due to the easy-axis orientation, whereby ***n*** is already parallel to $\mu\boldsymbol{x}$, in contrast to the R_{th} signal, which increases only above $\boldsymbol{H}_c$, when a sizeable field-induced moment ***m*** along $\mu\boldsymbol{x}$ appears. This observation confirms the high-field measurements in the previous geometry: a large $G_{\boldsymbol{n}}$ exists under spin bias, even without a strong reduction of the magnon gap. For ***H*** along ***x*** (Fig. 3a), the sharp drop in R_{el} at larger fields reflects the reorientation of ***n*** along the wire (***y***), which is perpendicular to $\mu\boldsymbol{x}$. This field-free spin transport demonstrates the potential of antiferromagnets for spintronic applications.

We next apply the field along ***y***. In Fig. 3b, R_{th} remains zero as the field is applied parallel to the platinum wires, while R_{el} is again non-zero, because the Néel order is insensitive to small fields. Thus, R_{el} reveals very distinct field dependences for the two device geometries. This finding shows that active control of the direction of the Néel order is key for spin transport, which is likely to be problematic in multi-domain thin films, explaining the low-efficiency spin transport of previous studies (with our single crystal, mono-domain devices can be considered; see Methods). The signal amplitude at small fields is comparable to the signal at the reduced magnon gap, possibly arising from different interface transmissivities (Methods). Another intrinsic contribution to the field dependence conceivably comes from spin-relaxation processes; at $\boldsymbol{H}_c$, the changes in the magnon dispersion curve could lead to different dissipation channels.

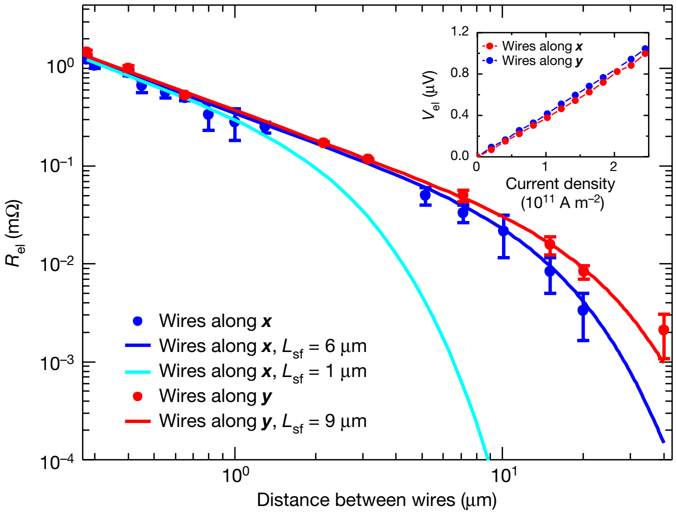


Fig. 4 | Dependence of the spin signal R_{el} on distance at 200 K. A signal persists for tens of micrometres at both $H_c = 6$ T (platinum wires along ***x***, blue) and low fields ($H_x = 2$ T; platinum wires along ***y***, red). Only R_{el} is shown, and is fitted (solid lines) by a one-dimensional spin diffusion equation[1] to extract the spin diffusion length L_{sf}. (R_{th} persists over more than 80 μm, but is difficult to appraise because the heating and the spin-current source are delocalized[32].) Error bars are a combination of the standard error of the mean and the measurement noise floor. Inset, the spin-bias voltage V_{el} is linear in both cases, as per a non-equilibrium spin-transport mechanism.

Finally, measuring spin diffusion lengths is key to determining the spin-transport regime, for which different predictions have been made (diffusive[15], quasi-ballistic[12] or spin-superfluid[25,26]). We focus in Fig. 4 on only the distance dependence of R_{el}. Figure 4 illustrates spin transport at 200 K over tens of micrometres, with a linear decay to a few micrometres in both geometries. This feature (alongside the R_{th} signal up to 80 μm; see Methods) might lead to the conclusion of a spin-superfluid regime[25,26], as discussed recently for low temperatures[19], even though what is observed here is certainly diffusive transport. Only for distances larger than the spin diffusion lengths is an exponential amplitude decay, predicating diffusive transport, observed[1]. Moreover, R_{el} varies linearly with the bias current and presents no threshold, which would exist owing to anisotropy effects in the spin-superfluid regime.

We determine the spin diffusion lengths to be 6 ± 1 μm and 9 ± 2 μm at $\boldsymbol{H}_c$ and at small applied fields, respectively, more than two orders of magnitude larger than those reported for antiferromagnetic–ferromagnetic multilayers[9,10]. In these systems, magnetic correlations between ferromagnetic and antiferromagnetic insulator grains are crucial[27]; ***n*** parallel to ***μ*** is critical for efficient spin transport. Another key issue in insulating antiferromagnetic thin films is the possibility of multiple domains affecting the mode polarization[9,27]. In accordance with theoretical work[28], the micrometre-scale spin diffusion lengths that we observed rely on the low Gilbert damping of α-Fe_2O_3 (of less than 10^{-4})[18,29], although the antiferromagnetic spin-wave relaxation processes involved remains an open debate[30,31]. This complexity is illustrated by the reduction in the spin diffusion length at $\boldsymbol{H}_c$ (9 ± 2 μm at small applied fields compared to 6 ± 1 μm), which indicates that the reduction in the magnon gap at $\boldsymbol{H}_c$ comes with stronger dissipation channels. Investigations of antiferromagnetic insulators with lower damping would reveal the unexplored relaxation processes of antiferromagnetic spin waves and open up the possibility of terahertz-frequency magnon spin-logic devices.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0490-7

Received: 27 April 2018; Accepted: 25 July 2018;
Published online 12 September 2018.

1. Cornelissen, L. J., Liu, J., Duine, R. A., Youssef, J. B. & van Wees, B. J. Long-distance transport of magnon spin information in a magnetic insulator at room temperature. *Nat. Phys.* **11**, 1022–1026 (2015).
2. Wadley, P. et al. Electrical switching of an antiferromagnet. *Science* **351**, 587–590 (2016).
3. Baierl, S. et al. Terahertz-driven nonlinear spin response of antiferromagnetic nickel oxide. *Phys. Rev. Lett.* **117**, 197201 (2016).
4. Jungwirth, T., Marti, X., Wadley, P. & Wunderlich, J. Antiferromagnetic spintronics. *Nat. Nanotechnol.* **11**, 231–241 (2016).
5. Baltz, V. et al. Antiferromagnetic spintronics. *Rev. Mod. Phys.* **90**, 015005 (2018).
6. Železný, J., Wadley, P., Olejník, K., Hoffmann, A. & Ohno, H. Spin transport and spin torque in antiferromagnetic devices. *Nat. Phys.* **14**, 220–228 (2018).
7. Jungwirth, T. et al. The multiple directions of antiferromagnetic spintronics. *Nat. Phys.* **14**, 200–203 (2018).
8. Lin, W., Chen, K., Zhang, S. & Chien, C. L. Enhancement of thermally injected spin current through an antiferromagnetic insulator. *Phys. Rev. Lett.* **116**, 186601 (2016).
9. Wang, H., Du, C., Hammel, P. C. & Yang, F. Spin transport in antiferromagnetic insulators mediated by magnetic correlations. *Phys. Rev. B* **91**, 220410 (2015).
10. Hahn, C. et al. Conduction of spin currents through insulating antiferromagnetic oxides. *Europhys. Lett.* **108**, 57005 (2014).
11. Morin, F. J. Electrical properties of α-Fe_2O_3 containing titanium. *Phys. Rev.* **83**, 1005–1010 (1951).
12. Bender, S. A., Skarsvåg, H., Brataas, A. & Duine, R. A. Enhanced spin conductance of a thin-film insulating antiferromagnet. *Phys. Rev. Lett.* **119**, 056804 (2017).
13. Chumak, A. V., Vasyuchka, V. I., Serga, A. A. & Hillebrands, B. Magnon spintronics. *Nat. Phys.* **11**, 453–461 (2015).
14. Keffer, F. & Kittel, C. Theory of antiferromagnetic resonance. *Phys. Rev.* **85**, 329–337 (1952).
15. Rezende, S. M., Rodríguez-Suárez, R. L. & Azevedo, A. Diffusive magnonic spin transport in antiferromagnetic insulators. *Phys. Rev. B* **93**, 054412 (2016).
16. Wu, S. M. et al. Antiferromagnetic spin Seebeck effect. *Phys. Rev. Lett.* **116**, 097204 (2016).
17. Seki, S. et al. Thermal generation of spin current in an antiferromagnet. *Phys. Rev. Lett.* **115**, 266601 (2015).
18. Elliston, P. R. & Troup, G. J. Some antiferromagnetic resonance measurements in α-Fe_2O_3. *J. Phys. C* **1**, 169–178 (1968).
19. Yuan, W. et al. Experimental signatures of spin superfluid ground state in canted antiferromagnet Cr_2O_3 via nonlocal spin transport. *Sci. Adv.* **4**, eaat1098 (2018).
20. Ganzhorn, K. et al. Temperature dependence of the non-local spin Seebeck effect in YIG/Pt nanostructures. *AIP Adv.* **7**, 085102 (2017).
21. Shan, J. et al. Criteria for accurate determination of the magnon relaxation length from the nonlocal spin Seebeck effect. *Phys. Rev. B* **96**, 184427 (2017).
22. Sinova, J., Valenzuela, S. O., Wunderlich, J., Back, C. H. & Jungwirth, T. Spin Hall effects. *Rev. Mod. Phys.* **87**, 1213–1260 (2015).
23. Zhang, S. S.-L. & Zhang, S. Magnon mediated electric current drag across a ferromagnetic insulator layer. *Phys. Rev. Lett.* **109**, 096603 (2012).
24. Wright, K. Focus: a trio of magnon transistors. *Physics* **11**, 23 (2018).
25. Qaiumzadeh, A., Skarsvåg, H., Holmqvist, C. & Brataas, A. Spin superfluidity in biaxial antiferromagnetic insulators. *Phys. Rev. Lett.* **118**, 137201 (2017).
26. Takei, S., Halperin, B. I., Yacoby, A. & Tserkovnyak, Y. Superfluid spin transport through antiferromagnetic insulators. *Phys. Rev. B* **90**, 094408 (2014).
27. Stiles, M. D. & McMichael, R. D. Model for exchange bias in polycrystalline ferromagnet-antiferromagnet bilayers. *Phys. Rev. B* **59**, 3722–3733 (1999).
28. Cramer, J. et al. Spin transport across antiferromagnets induced by the spin Seebeck effect. *J. Phys. D* **51**, 144004 (2018).
29. Sulymenko, O. R. et al. Terahertz-frequency spin hall auto-oscillator based on a canted antiferromagnet. *Phys. Rev. Appl.* **8**, 064007 (2017).
30. Kotthaus, J. P. & Jaccarino, V. Antiferromagnetic-resonance linewidths in MnF_2. *Phys. Rev. Lett.* **28**, 1649–1652 (1972).
31. Rezende, S. M. & White, R. M. Multimagnon theory of antiferromagnetic resonance relaxation. *Phys. Rev. B* **14**, 2939–2955 (1976).
32. Thiery, N. et al. Electrical properties of epitaxial yttrium iron garnet ultrathin films at high temperatures. *Phys. Rev. B* **97**, 064422 (2018).

Acknowledgements R.L. acknowledges the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement FAST number 752195. A.R., J.C. and M.K. acknowledge support from the Graduate School of Excellence Materials Science in Mainz (DFG/GSC 266). All authors from Mainz also acknowledge support from both MaHoJeRo (DAAD Spintronics network, project number 57334897) and SPIN+X (DFG SFB TRR 173). A.Q. and A.B. acknowledge support from the European Research Council via Advanced Grant number 669442 'Insulatronics'. S.A.B. and R.A.D. acknowledge support from Stichting voor Fundamenteel Onderzoek der Materie (FOM) and the European Research Council via Consolidator Grant number 725509 'SPINBEYOND'. A.Q., R.A.D., M.K. and A.B. were supported by the Research Council of Norway through its Centres of Excellence funding scheme, project number 262633 'QuSpin'.

Reviewer information *Nature* thanks T. Dietl, L. Morellón and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions R.L. and M.K. proposed and supervised the project. R.L. performed the experiments with A.R. and with technical support from J.C. and L.B. A.R. patterned the samples. R.L., S.A.B. and A.R. analysed the data with input from M.K., A.B. and R.A.D. S.A.B. performed the theoretical calculations with the assistance of A.Q., A.B. and R.A.D. R.L wrote the paper with S.A.B., A.R. and M.K. All authors commented on the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0490-7.

Reprints and permissions information is available at http://www.nature.com/reprints.

Correspondence and requests for materials should be addressed to R.L. or M.K.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

R-cut α-Fe$_2$O$_3$ single crystal and Morin transition. In Extended Data Fig. 1a, we represent the crystal structure of haematite (α-Fe$_2$O$_3$). Above the Morin temperature T_M the antiferromagnetic easy plane is the C plane (blue), whereas below it the antiferromagnetic easy axis is along the *c* axis (perpendicular to the C plane). The single crystal of α-Fe$_2$O$_3$ was obtained commercially from SurfaceNet and is orientated with $[11\bar{2}0]$ out of plane, that is, the *c* axis of the hexagonal structure is tilted about 33° below the surface plane. This orientation was chosen because of the large in-plane projection of the *c* axis in addition to the chemical stability of the sample terminating with the R plane (green plane in Extended Data Fig. 1a).

In Extended Data Fig. 1b, the abrupt magnetic transition from the weak ferromagnetic phase to the antiferromagnetic easy-axis phase in SQUID magnetometry occurs at T_M (260 K[11,33]). Below this temperature, we observe no residual magnetization, indicating the purity of the single crystal. Below T_M, $\boldsymbol{n}$ is parallel to the Dzyaloshinkii–Moriya interaction (DMI) vector, with no effect on the magnetic moments. Our sample maintains its insulating properties and exhibits no evidence of semiconducting behaviour from 300 K to 4 K. The Morin transition can be tuned by slightly doping α-Fe$_2$O$_3$ to shift T_M to higher (with Ir[34] or Sn[35] doping) or lower (with Ti or Ga doping[36]) temperatures.

Lithography and measurement procedure. The sample measured 5 mm × 5 mm × 0.5 mm, with a surface roughness of less than 0.1 nm. It was cleaned with acetone, isopropanol and deionized water before patterning. The samples were defined using electron-beam lithography and the subsequent deposition and lift-off of a 7-nm platinum layer was performed by d.c. sputtering in an argon atmosphere. The non-local wires were contacted using a bilayer of Cr(6 nm)/Au(32 nm). The geometry consists of three wires 50 μm in length (L) and 350 nm wide with wires patterned parallel and perpendicular to the sample easy axis. The wire separations differed, ranging from 200 nm to 80 μm. In Fig. 1b we show an SEM image of a typical device.

For separations of more than 1 μm, a second geometry (with double the wire length, $2L$) was used to limit the geometric effect on the signal. The injection current was also doubled to increase the signal-to-noise ratio. To allow for accurate scaling of the signal, calibration distances of 500 nm were added where the scaled signal is consistent with comparable distances in the first geometry.

The sample was mounted onto a piezo-rotating element in a variable-temperature insert installed in a superconducting magnet capable of 12 T. The field was swept in-plane either parallel or perpendicular to the antiferromagnetic easy axis.

The central wire was used to carry a direct charge current, with either positive or negative polarity, of 300 μA for devices with a length L (600 μA for $2L$ length). Owing to the SHE, a spin current flows towards the Pt–α-Fe$_2$O$_3$ interface. The measured voltage was detected in the left and right detector wires using a nanovoltmeter, owing to the inverse SHE creating a charge flow. This voltage was recorded as a function of spatial separation, external field, applied bias current and angle between the charge current and the field. The voltage was recorded for positive and negative injection currents so that, as a function of time, the current traces a square-wave-like signal where a delay was added between reversing the current and the start of data collection to allow for any rise time in the current source and any capacitor-like effects in the nanostructures. The thermal transport signal is independent of current direction whereas the SHE will cause a spin current with opposite polarization depending on the current direction, allowing for the separation of the two signals as detailed in the main text as the sum and difference of the two signals.

Domain size in the canted antiferromagnetic phase. Above the Morin temperature, we can image the canted antiferromagnetic domains by using MOKE microscopy. We observe domain sizes of hundreds of micrometres (Extended Data Fig. 2). The emergence of such domains (separated by 180° domain walls) through the Morin transition by the fusion of two antiferromagnetic domains (separated by 90° antiferromagnetic domain walls) has been reported previously[37]. Our findings similarly indicate that the pure antiferromagnetic domains are of micrometre sizes, and it is therefore reasonable to model the antiferromagnetic spin transport by assuming a single domain between our platinum wires (particularly at high field).

Modelling the Néel order and magnetization orientations. Hematite has a uniaxial anisotropy and a DMI, which are both rotationally invariant around a common axis $\boldsymbol{x}'$. Below T_M, the anisotropy is along the indicated easy axis, and at zero field any induced magnetization vanishes, with the Néel order oriented along the easy axis. We model the macrospin of the α-Fe$_2$O$_3$ crystal using a two spin model, with energy

$$E = J\boldsymbol{m}_1\cdot\boldsymbol{m}_2 - \frac{K}{2}[(\boldsymbol{m}_1\cdot\boldsymbol{x}')^2 + (\boldsymbol{m}_2\cdot\boldsymbol{x}')^2] + \frac{1}{2}\boldsymbol{D}\cdot\boldsymbol{m}_1\times\boldsymbol{m}_2 - \boldsymbol{H}\cdot(\boldsymbol{m}_1+\boldsymbol{m}_2) \quad (2)$$

where $\boldsymbol{m}_1$ and $\boldsymbol{m}_2$ are unit vectors that represent the magnetic orientation of each sublattice, K is the anisotropy, J is the exchange, $\boldsymbol{D} = D\boldsymbol{x}'$ is the DMI vector and $\boldsymbol{H}$ is the applied field, all with units of energy.

At zero field, $\boldsymbol{m} = \boldsymbol{0}$ whereas $\boldsymbol{n}$ lies along $\boldsymbol{x}'$. When a field $\boldsymbol{H}$ is applied along the $\boldsymbol{x}'$ direction, the α-Fe$_2$O$_3$ crystal is rotationally invariant and admits a simple analytic solution. Below $H_c = \sqrt{2JK}$, $\boldsymbol{n}$ lies along the $\boldsymbol{x}'$ axis and the magnetization vanishes; however, when $H > H_c$, α-Fe$_2$O$_3$ is in a spin-flop phase, with a paramagnetic moment $\boldsymbol{m}$ induced along the $\boldsymbol{x}'$ direction and $\boldsymbol{n}$ perpendicular to $\boldsymbol{m}$ and $\boldsymbol{H}$. To obtain the magnetic state, we numerically minimize the energy E in equation (2) with respect to $\boldsymbol{n}$ and $\boldsymbol{m}$ by solving the Landau–Lifshitz–Gilbert equations.

Denoting the sample plane as the x–y plane, the easy axis $\boldsymbol{x}'$ makes an approximately 33° angle with the $\boldsymbol{x}$ axis (Extended Data Fig. 3a). We find that when the field $\boldsymbol{H}$ is applied in the $\boldsymbol{y}$ direction (perpendicular to the easy axis $\boldsymbol{x}'$), a paramagnetic moment $\boldsymbol{m}$ is induced along the field. However, the presence of the DMI energy (proportional to $\boldsymbol{D}\cdot\boldsymbol{m}\times\boldsymbol{n}$) complicates the situation. The DMI energy is minimized when $\boldsymbol{m}\times\boldsymbol{n}$ is collinear with $\boldsymbol{x}'$. Thus, $\boldsymbol{n}$, which is initially in the $\boldsymbol{x}'$ direction, rotates at large fields (around $2\boldsymbol{H}_c$ in the literature[38]) nearly out of the plane in the x–z plane until it is perpendicular to $\boldsymbol{x}'$ (Extended Data Fig. 3b). Note that for an R-plane crystal, the direction of $\boldsymbol{n}$ is easily controllable and well defined above $\boldsymbol{H}_c$ (by the intersection between the plane perpendicular to the easy axis and the plane perpendicular to the applied field), which is not achieved as easily for C- and A-plane surfaces by applying an in-plane field.

When the field is applied in the $\boldsymbol{x}$ direction, typical spin-flop behaviour is observed. At small fields $\boldsymbol{n}$ cants towards the $\boldsymbol{z}$ axis, while a paramagnetic moment $\boldsymbol{m}$ is induced in the x–z plane (case (1) in Extended Data Fig. 3c); thus, at small fields, $\boldsymbol{m}$ and $\boldsymbol{n}$ lie in the x–z plane. When the applied field is sufficiently large (around $\boldsymbol{H}_c$), $\boldsymbol{n}$ gradually reorients towards $\boldsymbol{y}$ (case (2) in Extended Data Fig. 3c).

Theoretical model of spin Hall magnetoresistance. The local magnetic orientations of α-Fe$_2$O$_3$ in response to an external field $\boldsymbol{H}$ manifest in the spin-Hall magnetoresistance (SMR) of the local platinum wires. To show that the numerical orientations of $\boldsymbol{n}$ and $\boldsymbol{m}$ obtained above are consistent with the experimental SMR measurements (Methods section 'Experimental and theoretical spin Hall magnetoresistance'), we briefly discuss here a model for the dependence of SMR on $\boldsymbol{n}$ and $\boldsymbol{m}$.

The following is an expression for the interfacial spin current $\boldsymbol{j}_s$ based on symmetry arguments:

$$\boldsymbol{j}_s = \frac{g_{\boldsymbol{n}}^{(\uparrow\downarrow)}}{4\pi}\boldsymbol{n}\times(\boldsymbol{\mu}\times\boldsymbol{n}) + \frac{g_{\mathrm{r}\boldsymbol{m}}^{(\uparrow\downarrow)}}{4\pi}\boldsymbol{m}\times(\boldsymbol{\mu}\times\boldsymbol{m}) + \frac{g_{\mathrm{i}\boldsymbol{m}}^{(\uparrow\downarrow)}}{4\pi}\boldsymbol{\mu}\times\boldsymbol{m} + \boldsymbol{j}_s^{(\mathrm{T})} \quad (3)$$

where $g_{\boldsymbol{n}}^{(\uparrow\downarrow)}$ is the Néel spin-mixing conductance, and $g_{\mathrm{r}\boldsymbol{m}}^{(\uparrow\downarrow)}$ and $g_{\mathrm{i}\boldsymbol{m}}^{(\uparrow\downarrow)}$ are the real and imaginary parts of the magnetic spin-mixing conductance. The assumption of no sublattice-symmetry breaking at the interface allows us to drop all terms that are linear in $\boldsymbol{n}$. The first three terms originate from the static orientations of $\boldsymbol{n}$ and $\boldsymbol{m}$, and $\boldsymbol{j}_s^{(\mathrm{T})}$ is the spin current that arises from thermal spin waves. This interfacial spin current describes the flow of spins between heavy-metal electron spins and incoherent antiferromagnetic insulator spin waves. Hence, it facilitates the injection and detection of thermal spin-wave currents, and ultimately gives rise to equation (1). In theoretical treatments of ferromagnetic-insulator–heavy-metal bilayers, SMR is attributed to the static order parameter $\boldsymbol{n}$. Therefore, we similarly neglect the role of $\boldsymbol{j}_s^{(\mathrm{T})}$ in SMR.

The expression for the spin current being absorbed by the heavy metal is

$$\boldsymbol{j}_s = -\frac{\hbar}{2e^2}\sigma_0(\partial_{\boldsymbol{z}}\boldsymbol{\mu} - \theta_{\mathrm{SH}}\boldsymbol{z}\times\boldsymbol{E}) \quad (4)$$

where $\boldsymbol{E}$ is the applied electric field inside the heavy-metal layer (assumed to be uniform in the $\boldsymbol{z}$ direction), σ_0 is the bare conductance of the wire, θ_{SH} is the spin Hall angle, e is the electron charge, $\hbar$ is the reduced Planck constant and the right-hand side is evaluated at the interface.

When $g_{\mathrm{r}\boldsymbol{m}}^{(\uparrow\downarrow)} = g_{\mathrm{i}\boldsymbol{m}}^{(\uparrow\downarrow)} = 0$, the expression for $\boldsymbol{j}_s$ is identical to that obtained for a ferromagnet with a real spin-mixing conductance, with the Néel order replacing the magnetization. Last, the charge current $\boldsymbol{j}$ that flows through a wire is

$$\boldsymbol{j} = \sigma_0\boldsymbol{E} - \frac{\sigma_{\mathrm{SH}}}{e}\boldsymbol{\nabla}\times\boldsymbol{\mu} \quad (5)$$

The general form of $\boldsymbol{\mu}$ inside the heavy-metal wire is obtained from a spin diffusion equation: $\boldsymbol{\mu} = \boldsymbol{A}\mathrm{e}^{-z/\lambda} + \boldsymbol{B}\mathrm{e}^{z/\lambda}$, where λ is the heavy-metal spin-flip length. The parameters $\boldsymbol{A}$ and $\boldsymbol{B}$ and thus $\boldsymbol{\mu}$ are obtained by equating equations (3) and (5). Averaging over the thickness of the heavy-metal wires in the $\boldsymbol{z}$ direction and inverting to find $\boldsymbol{E} = \rho\boldsymbol{j}_{\mathrm{ave}}$ (where $\boldsymbol{j}_{\mathrm{ave}}$ is the thickness-averaged current) we obtain the 2×2 resistivity tensor ρ, which depends on $\boldsymbol{n}$ and $\boldsymbol{m}$ through $\boldsymbol{A}$ and $\boldsymbol{B}$.

Experimental and theoretical spin Hall magnetoresistance. *Platinum wires along* $\boldsymbol{x}$. In the injector wire, we measure (with a two-point detection scheme) the field-induced SMR for $\boldsymbol{H}$ parallel (along $\boldsymbol{x}$) or perpendicular (along $\boldsymbol{y}$) to the wire, as seen in Extended Data Fig. 4. In this configuration, the easy axis of α-Fe$_2$O$_3$ has a projection along the wire and no projection perpendicular to it at zero applied field.

From the local SMR measurements, we detect the presence of a spin-flop at around 6 T when the field is applied along $\boldsymbol{y}$, making a 33° angle between the applied field and the easy axis. From this measurement, we extract an SMR ratio of $\Delta R/R = 2.5 \times 10^{-4}$ at $\boldsymbol{H}_c$. With the theoretical formalism presented previously, we fit the SMR curves for the two field orientations, reproducing qualitatively both the amplitude of the signals and the magnetic field reorientation at $\boldsymbol{H}_c$.

Platinum wires along $\boldsymbol{y}$. The SMR for platinum wires along $\boldsymbol{y}$ is shown in Extended Data Fig. 5 for fields parallel and perpendicular to the wire. From these curves, we again extract the SMR ratio, obtaining a value of $\Delta R/R = 4 \times 10^{-4}$ at $\boldsymbol{H}_c$, about twice the value for the previous orientation. Similarly, the fitted values of the spin-mixing conductances are larger for this device geometry by a factor of about two. Being careful in making conclusions about the origins of these differences (which could include lithographic effects, interface transmissivities or thermal spin waves), we note it is in agreement with the larger intrinsic spin-transport signals for this orientation (Fig. 3, Methods section 'Model for non-local spin-transport signal').

Furthermore, for the $\boldsymbol{x}$ scan the experimental and theoretical values of $\boldsymbol{H}_c$ are both around 6 T, whereas for the $\boldsymbol{y}$ scan ($\boldsymbol{H}$ perpendicular to the easy axis) the theoretical DMI-induced reorientation (Extended Data Fig. 3c) shifts the experimental value slightly (8 T instead of 10 T). This discrepancy arises from the difficulty of simultaneously fitting these two reorientations (classical $\boldsymbol{H}_c$ and DMI-induced reorientation) by considering only the easy axis and DMI anisotropies.

Model for non-local spin-transport signal. Equation (1) is a minimal model for two-channel transport by $\boldsymbol{m}$ and $\boldsymbol{n}$. The factors of $(\boldsymbol{n} \cdot \boldsymbol{e}_\perp)^2$ and $(\boldsymbol{m} \cdot \boldsymbol{e}_\perp)^2$ arise from the spin injection by $\boldsymbol{\mu} \approx \mu \boldsymbol{e}_\perp$ into the antiferromagnetic insulator Néel order or magnetization and the subsequent spin injection into the detector. By contrast, the spin accumulation in the injector wire has no role in the spin-Seebeck signal, hence $\boldsymbol{n} \cdot \boldsymbol{e}_\perp$ and $\boldsymbol{m} \cdot \boldsymbol{e}_\perp$ in equation (1). All four of the coefficients depend generally on $\boldsymbol{n}$ and $\boldsymbol{m}$, so sublattice symmetry of the antiferromagnetic insulator implies $S_{\boldsymbol{n}}(\boldsymbol{n}) = -S_{\boldsymbol{n}}(-\boldsymbol{n})$, and so on. Finally, although R_{el}, $G_{\boldsymbol{n}}$ and $G_{\boldsymbol{m}}$ are measured in units of V A^{-1}, they are measures of the spin conductance of the antiferromagnetic insulator. Because the spin-Seebeck signal is quadratic in current, R_{th} and therefore $S_{\boldsymbol{n}}$ and $S_{\boldsymbol{m}}$ are measured in units of V A^{-2} but represent a measure of the spin-Seebeck conductance.

Then, equation (1) depends on $\boldsymbol{H}$ via the terms $(\boldsymbol{n} \cdot \boldsymbol{e}_\perp)$, $(\boldsymbol{n} \cdot \boldsymbol{e}_\perp)^2$, $(\boldsymbol{m} \cdot \boldsymbol{e}_\perp)$ and $(\boldsymbol{m} \cdot \boldsymbol{e}_\perp)^2$. From the determination of $\boldsymbol{m}$ and $\boldsymbol{n}$ (Methods section 'Modelling the Néel order and magnetization orientations'), we make the following predictions about the non-local spin signals.

For wires oriented along $\boldsymbol{x}$, $\boldsymbol{e}_\perp = \boldsymbol{y}$ and $\boldsymbol{\mu} \approx \mu \boldsymbol{y}$. When $\boldsymbol{H}$ is applied in the $\boldsymbol{y}$ direction, $\boldsymbol{n}$ lies in the x–z plane and $\boldsymbol{m}$ along $\boldsymbol{y}$ (Extended Data Fig. 3b). Thus, the non-local signal cannot arise from the Néel order because $\boldsymbol{n} \cdot \boldsymbol{y} = (\boldsymbol{n} \cdot \boldsymbol{y})^2 = 0$. Instead, the non-local signals must arise from the paramagnetization $\boldsymbol{m}$, but $\boldsymbol{m}$ vanishes at small fields. We therefore expect an enhanced signal with applied field (Fig. 2b). At small fields applied in the $\boldsymbol{x}$ direction, $\boldsymbol{n}$ and $\boldsymbol{m}$ both lie in the x–z plane (case (1) in Extended Data Fig. 3c), so the signal should vanish. As the field is increased, the anisotropy and DMI fields reorient $\boldsymbol{n}$ along $\boldsymbol{y}$ (case (2) in Extended Data Fig. 3c), opening up the possibility of a signal at higher fields arising from the Néel order (Fig. 2a).

For wires oriented along $\boldsymbol{y}$, $\boldsymbol{e}_\perp = \boldsymbol{x}$ and $\boldsymbol{\mu} \approx \mu \boldsymbol{x}$. When the field is applied in the $\boldsymbol{x}$ direction, we find that for small fields $\boldsymbol{n} \cdot \boldsymbol{x}$ is non-zero, giving rise to a non-local signal. At higher fields, $\boldsymbol{n}$ reorients owing to the DMI and the anisotropy fields and the signal is reduced (Fig. 3b). At large fields along $\boldsymbol{y}$ (about $2\boldsymbol{H}_c$), $\boldsymbol{n}$ reorients as a result of only the DMI.

Equation (1) also depends on $\boldsymbol{H}$ through the structural transport coefficients $G_{\boldsymbol{n}}$, $G_{\boldsymbol{m}}$, $S_{\boldsymbol{n}}$ and $S_{\boldsymbol{m}}$, which have unknown general dependences. However, the field dependence of these coefficients for fields applied along the easy axis is known from a stochastic Landau–Lifshitz phenomenology[12] and drift-diffusive treatment[15,39].

The spin conductance G is predicted to diverge as one of the gaps of the two spin-wave branches closes[12]. We therefore approximate the low-energy behaviour of G by

$$G \approx G_0 (H_c/\Delta_{\min}) \tag{6}$$

where G_0 is a constant to be determined by fitting the data and $\Delta_{\min}$ is the smaller of the two spin-wave branch gaps under a field $\boldsymbol{H}$. To obtain $\Delta_{\min}$ for a given field, we find the ground-state orientations $\boldsymbol{m}$ and $\boldsymbol{n}$ and then linearize the Landau–Lifshitz–Gilbert equations of motion for $\boldsymbol{m}_1$ and $\boldsymbol{m}_2$ around this state; we thereby find two eigenfrequencies, corresponding to the two main spin-wave branches. At zero field, these are the degenerate oscillations of the Néel order, with opposite helicities. At non-zero field, the degeneracy is lifted (dashed lines in Fig. 2b). The theoretical R_{el} is then determined by the inverse of the lowest magnon gap Δ and by the projection $\mu \boldsymbol{y} \cdot \boldsymbol{n}$.

When $G_{\boldsymbol{m}} = 0$ ($G \approx G_{\boldsymbol{n}}$), we find that equations (1) and (6) qualitatively reproduce the signal shown in Fig. 2a, and the other orientation, as shown in Extended Data Fig. 6. Our simple model wherein G_0 is field-independent is consistent with the prediction that the dominant contribution to the spin current is symmetric in $\boldsymbol{H}$.

The data are consistent with $S_{\boldsymbol{m}} \gg S_{\mathbf{n}}$ (see below), so we focus on $S_{\boldsymbol{m}}$. Because $S_{\boldsymbol{n}}$ is known to have a continuous dependence on the field[12,16,39] (unlike $G_{\boldsymbol{n}}$), we posit a continuous dependence of $S_{\boldsymbol{m}}$ as well. Taking the lowest-order term, we model $S_{\boldsymbol{m}} \approx S_0$ as field-independent because the dominant field dependence is captured by $\boldsymbol{m} \cdot \boldsymbol{e}_\perp$.

We show the fits of the data with our model, equation (6) and $S_{\boldsymbol{m}} \approx S_0$ in Extended Data Fig. 6. We obtain similar spin-Seebeck conductance coefficients ($S_0 \approx 60$ V A^{-2}) for the two device geometries; G_0 is much larger for platinum wires along $\boldsymbol{y}$ (Fig. 3) than along $\boldsymbol{x}$ (Fig. 2). This is in accordance with the SMR measurements presented in Methods section 'Experimental and theoretical spin Hall magnetoresistance' and fits as $G_0 \propto \left(g_{\mathbf{n}}^{(\uparrow\downarrow)}\right)^2$. Similarly to the SMR measurements, the difference in intrinsic spin conductance for the two geometries might arise from the patterning processes, or different transmissivities and anisotropic spin diffusion lengths, which are not taken into account in our model.

The only slight discrepancy between our theoretical model and the experiments concerns the high-field spin signal R_{el} shown in Extended Data Fig. 6d. In this case, the applied field is perpendicular to the easy axis and a spin reorientation happens only at large field owing to the presence of the DMI. In our SMR data this high-field reorientation occurs at 10 T (Extended Data Fig. 5), whereas in our SMR and non-local fits it occurs around 8 T for our chosen set of K, J and $\boldsymbol{D} = D\boldsymbol{x}'$. The fact that this reorientation occurs at larger fields in the experiment explains why we do not observe the same increase in the spin signal as our model predicts. Higher-order anisotropy terms could explain why we cannot obtain exactly this magnetic reorientation at its experimental value by using only three parameters (K, J and D). Another contribution to this discrepancy is that the magnetic dynamics change from circular to elliptic, and then to nearly linear at high fields[40] applied perpendicular to the easy axis, an effect that is not taken into account. In easy-axis antiferromagnets such as α-Fe$_2$O$_3$, the presence of circularly polarized magnon modes is an essential difference compared to the well studied family of easy-plane antiferromagnets such as the insulating NiO[9,10]. In the latter case, spin transport was reported over only nanometre-scale distances, not more than for metallic antiferromagnets with larger damping such as IrMn[28,41], which can be attributed to the easy-plane symmetry of the system, to the presence of disorder in antiferromagnetic thin films and to the inefficient spin injection in ferromagnetic–antiferromagnetic systems with non-collinear magnetic orders[9,41]. Finally, the spin transport and coupling can also be affected by phonons[42], which is not currently taken into account explicitly in our description. However, previous studies on non-magnetic SrTiO$_3$ have not observed any transport beyond the tunnelling range[9].

Spin conductance. To compare the spin conductance in ferromagnetic and antiferromagnetic insulators, we estimate the spin conductance of the antiferromagnet using the spin-resistor circuit model[1,43,44]:

$$\frac{R_{\mathrm{nl}}}{R_{\mathrm{SMR}}} = \frac{R^{\mathrm{s}}_{\mathrm{Pt}}}{2R^{\mathrm{s}}_{\mathrm{Pt}} + 2R^{\mathrm{s}}_{\mathrm{i}} + R^{\mathrm{s}}_{\alpha\cdot\mathrm{Fe_2O_3}}}$$

where R_{nl} is the non-local resistance, R_{SMR} is the change in resistance due to SMR, $R^{\mathrm{s}}_{\mathrm{Pt}}$ is the spin resistivity of the platinum wires, $R^{\mathrm{s}}_{\mathrm{i}}$ is the interfacial spin resistance and $R^{\mathrm{s}}_{\alpha\cdot\mathrm{Fe_2O_3}}$ is the spin conductance of hematite. For the shortest distance ($d \approx 300$ nm), we obtain a non-local-to-SMR ratio of $R_{\mathrm{nl}}/R_{\mathrm{SMR}} \approx 1/1{,}000$, about one order of magnitude smaller than obtained previously[1] ($R_{\mathrm{nl}}/R_{\mathrm{SMR}} \approx 1/57$ for $d \approx 300$ nm). However, the calculated spin conductance of platinum is higher in our experiment (1.3×10^{15} Ω^{-1} m^{-2} compared to 10^{14} Ω^{-1} m^{-2}). We obtain a similar spin resistance of about 0.02 Ω (with a Pt–α-Fe$_2$O$_3$ interface area of 1.75×10^{-11} m^{-2}) and an effective spin-mixing conductance of the Pt–α-Fe$_2$O$_3$ interface of[44,45] 10^{13} Ω^{-1} m^{-2}, which is consistent across both device geometries.

In a bulk crystal, extracting the spin conductivity requires some assumptions compared to thin films. First, the spin channel thickness t_c should theoretically be of the order of the spin diffusion length and not the sample thickness. This implies that no spins are dissipated into the crystal (which was shown not to be correct in the ferromagnetic case because the signal amplitude decreases with increasing film thickness[46]). If these two assumptions hold for our system, then we can extract a spin conductivity of about 10^5 S m^{-1} with channel dimensions of $t_c \approx L_{\mathrm{sf}} \approx 7$ μm, $w_c = 50$ μm and $L_c \approx 300$ nm, where w_c is the channel length defined by the wire length, L_{sf} is the spin diffusion length and L_c is the lateral size of the channel. From these first measurements, we ascertain that the antiferromagnetic spin conductivity is of the order of that for the ferromagnetic case. For a more quantitative analysis, samples of different thicknesses would need to be considered.

There are two further points that hamper a quantitative comparison. First, the platinum spin conductance varies by one order of magnitude in the literature (10^{14}–10^{15} Ω^{-1} m^{-2})[47,48]. Second, the interface properties depend strongly on the surface treatment in ferromagnets[48], in addition to our measurements being carried out at 200 K (compared to room temperature in previous measurements[1]). Future work is needed to investigate the interface effects and temperature dependence of diffusive transport in antiferromagnets.

Angular dependence. For devices with platinum wires along $\boldsymbol{x}$, we investigate the symmetry of the two spin-transport signals by rotating the sample under a fixed field. In Extended Data Fig. 7a, b, the spin transport and the thermal signals are shifted 90° relative to each other. There, $\alpha = 0°$ ($\alpha = 90°$) corresponds to $\boldsymbol{H}$ parallel to $\boldsymbol{\mu}_0$ ($\boldsymbol{H}$ parallel to the easy axis), where the spin signal is minimum (maximum) and the thermal signal is maximal (minimum). This confirms that the different origins of the two signals is due to either $G_{\boldsymbol{n}}$ or $S_{\boldsymbol{m}}$. R_{el} is present for $\boldsymbol{n}$ along $\boldsymbol{\mu}$ and is maximum at $\boldsymbol{H}_{\mathrm{c}}$ (around 6 T; Extended Data Fig. 7a), whereas the thermal signal is present for $\boldsymbol{m}$ along $\boldsymbol{\mu}$ and is thus linear with the amplitude of the applied field (Extended Data Fig. 7b). The spin injection signal has a $\sin^2(\alpha)$ dependence because both the injecting (SHE) and detecting (inverse SHE) mechanisms are α-dependent and the signal follows $\boldsymbol{n}$ rather than $\boldsymbol{m}$ as in ferromagnets.

R_{th} exhibits a $\cos(\alpha)$ dependence from the inverse-SHE detection but maintains the same symmetry as in ferromagnets owing to the similar transport mechanism through $\boldsymbol{m}$.

Origin and distance dependence of the spin-Seebeck signal. R_{th} has a similar distance dependence for the two device geometries (Extended Data Fig. 8), which illustrates that, far above $\boldsymbol{H}_{\mathrm{c}}$, the magnon dispersion is similar for these two configurations.

The R_{th} signal can be detected up to a distance of 80 μm, with a slower decay compared to the direct spin-injected signal. Up to 40 μm, the decay of the signal is linear, which could be associated with spin-superfluidity[19], whereas the symmetry of the system (easy-axis and not easy-plane antiferromagnet) and the elevated temperature (200 K) allow us to exclude this regime (in addition to the features of the spin-injected signal discussed in the main text).

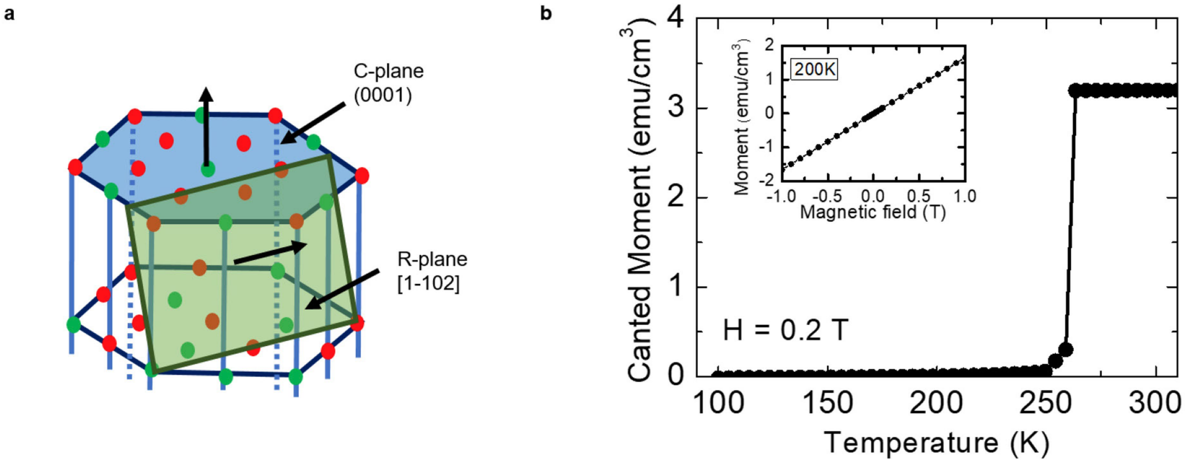
Furthermore, at low temperatures, the spin signal disappears (filled and open blue symbols in Extended Data Fig. 9), as expected for a diffusive rather than spin-superfluid regime, which anticipates a large increase as the anisotropy is compensated by either a magnetic field at $\boldsymbol{H}_{\mathrm{c}}$ or reduced temperature. In parallel, the R_{th} signal persists (filled and open black symbols in Extended Data Fig. 9), as in a previous claim[19] of spin-superfluid transport in thin films of $\mathrm{Cr_2O_3}$, based on only thermal signals at low temperature. Nevertheless, contrary to this claim, the absence of a spin-injected signal is not associated with poor platinum–antiferromagnetic-insulator interfaces and low spin-mixing conductances but is due only to its diffusive nature, because we observe a clear spin-injected transport signal at higher temperature. A reduction in the spin-injected signal while maintaining a thermal signal is an effect that has been seen previously in temperature-dependent measurements on ferromagnets[20,49]. This discrepancy between spin-injected and spin-Seebeck signals most probably arises from complex contributions related to the thermal source.

Although the transport regime can easily be analysed from the spin-injected signal (owing to the localization of the SHE-induced spin accumulation in the injector), much care should be taken in analysing the spin-Seebeck signal, as discussed in several recent works[20,21,46,50]. For example, the fact that the heating source is not localized to a single point and can generate thermal magnons far from the platinum injector should be taken into account[21,32]. At low temperatures, the Joule heating (with a current density of around 2.8×10^{11} A m^{-2} as used in Extended Data Fig. 9, similar to previous work[19]) also generates a non-negligible increase in temperature in the injector (of around 20 K for a helium bath of a few kelvin), which keeps the injector in the temperature region of the phonon peak, as demonstrated in Extended Data Fig. 10. In the entire temperature range, spatially extended thermal gradients are also generated by Joule heating. The non-local thermal signal can also include contributions from thermal magnons generated far from the injector. Furthermore, the spin-Seebeck signal can be driven by a heat-flow-affected regime when the condition $t_{\mathrm{s}} \ll \lambda$ is not fulfilled (with t_{s} the sample thickness and λ the spin diffusion length) as shown previously[21]. In this case, the spin-Seebeck signal is dominated by heat backflow in the detector for long distances. Contrary to ferromagnets, this criterion can also be true in ultrathin antiferromagnetic films for which the effective spin diffusion can be a few nanometres owing to the presence of antiferromagnetic domains[9,10,41,51], which would also reduce exponentially any spin-superfluid state[52]. Finally, as seen in the main text, R_{th} arises for the field-induced magnetization. Above $\boldsymbol{H}_{\mathrm{c}}$, the magnetic moments orient perpendicular to the magnetic field and the spins cant along the magnetic field. This increases the net magnetization in the direction of the magnetic field and leads to an increase in the magnetic anisotropy, which would destroy any spin-superfluid state that was produced regardless of its origin[25].

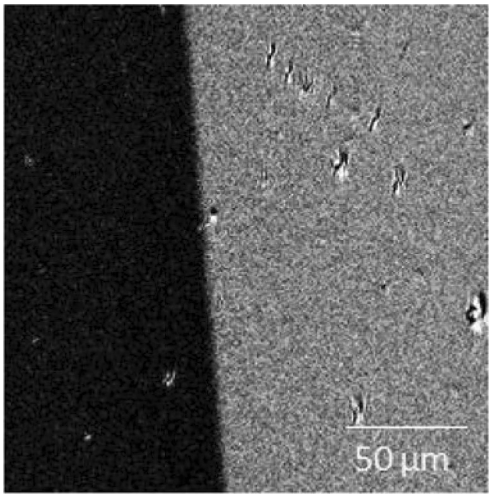
To conclude, as shown for our samples, caution must be taken when determining the spin-transport regime from only the R_{th} signal, particularly in antiferromagnetic samples where this signal can be misleading.

Data availability. The data that support the findings of this study are available from the corresponding authors on reasonable request.

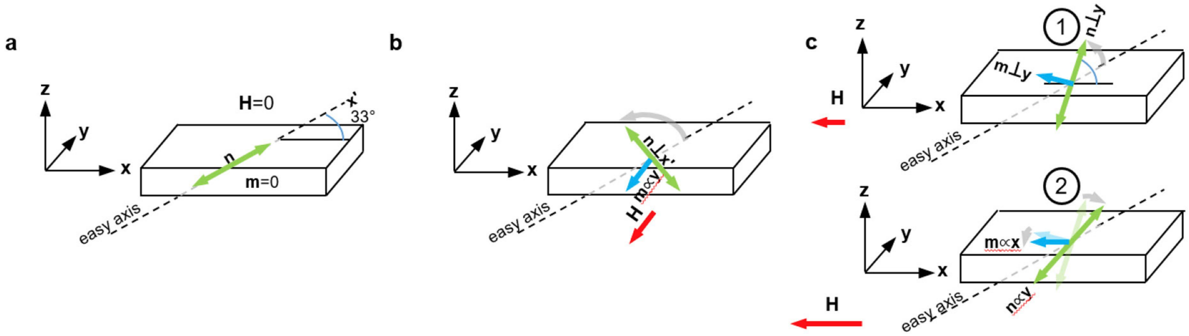
33. Moriya, T. Anisotropic superexchange interaction and weak ferromagnetism. *Phys. Rev.* **120**, 91–98 (1960).
34. Liu, J. Z. Morin transition in hematite doped with iridium ions. *J. Magn. Magn. Mater.* **54–57**, 901–902 (1986).
35. Ellis, D. S. et al. Magnetic states at the surface of α-$\mathrm{Fe_2O_3}$ thin films doped with Ti, Zn, or Sn. *Phys. Rev. B* **96**, 094426 (2017).
36. Besser, P. J., Morrish, A. H. & Searle, C. W. Magnetocrystalline anisotropy of pure and doped hematite. *Phys. Rev.* **153**, 632–640 (1967).
37. Mitsek, A. I. & Gaidanskii, P. F. The influence of domain structure on the magnetic properties of hematite. *Phys. Status Solidi a* **4**, 319–326 (1971).
38. Morrison, B. R., Morrish, A. H. & Troup, G. J. High-field antiferromagnetic resonance in α-$\mathrm{Fe_2O_3}$. *Phys. Status Solidi b* **56**, 183–195 (1973).
39. Rezende, S. M., Rodríguez-Suárez, R. L. & Azevedo, A. Theory of the spin Seebeck effect in antiferromagnets. *Phys. Rev. B* **93**, 014425 (2016).
40. Gurevich, A. G. & Melkov, G. A. *Magnetization Oscillations and Waves* **Ch. 3**, 69–90 (CRC Press, Boca Raton, 1996).
41. Jungfleisch, M. B., Zhang, W. & Hoffmann, A. Perspectives of antiferromagnetic spintronics. *Phys. Lett. A* **382**, 865–871 (2018).
42. Korenev, V. L. et al. Long-range *p*–*d* exchange interaction in a ferromagnet–semiconductor hybrid structure. *Nat. Phys.* **12**, 85–91 (2016).
43. Zhang, S. S.-L. & Zhang, S. Spin convertance at magnetic interfaces. *Phys. Rev. B* **86**, 214424 (2012).
44. Cornelissen, L. J., Peters, K. J. H., Bauer, G. E. W., Duine, R. A. & van Wees, B. J. Magnon spin transport driven by the magnon chemical potential in a magnetic insulator. *Phys. Rev. B* **94**, 014412 (2016).
45. Chen, Y.-T. et al. Theory of spin Hall magnetoresistance. *Phys. Rev. B* **87**, 144411 (2013).
46. Shan, J. et al. Influence of yttrium iron garnet thickness and heater opacity on the nonlocal transport of electrically and thermally excited magnons. *Phys. Rev. B* **94**, 174437 (2016).
47. Castel, V., Vlietstra, N., Ben Youssef, J. & van Wees, B. J. Platinum thickness dependence of the inverse spin-Hall voltage from spin pumping in a hybrid yttrium iron garnet/platinum system. *Appl. Phys. Lett.* **101**, 132414 (2012).
48. Jungfleisch, M. B., Lauer, V., Neb, R., Chumak, A. V. & Hillebrands, B. Improvement of the yttrium iron garnet/platinum interface for spin pumping-based applications. *Appl. Phys. Lett.* **103**, 022411 (2013).
49. Cornelissen, L. J. et al. Nonlocal magnon-polaron transport in yttrium iron garnet. *Phys. Rev. B* **96**, 104441 (2017).
50. Zhou, X. J. et al. Lateral transport properties of thermally excited magnons in yttrium iron garnet films. *Appl. Phys. Lett.* **110**, 062407 (2017).
51. Wang, H., Du, C., Hammel, P. C. & Yang, F. Antiferromagnonic spin transport from $\mathrm{Y_3Fe_5O_{12}}$ into NiO. *Phys. Rev. Lett.* **113**, 097202 (2014).
52. Upadhyaya, P., Kim, S. K. & Tserkovnyak, Y. Magnetic domain wall floating on a spin superfluid. *Phys. Rev. Lett.* **118**, 097201 (2017).



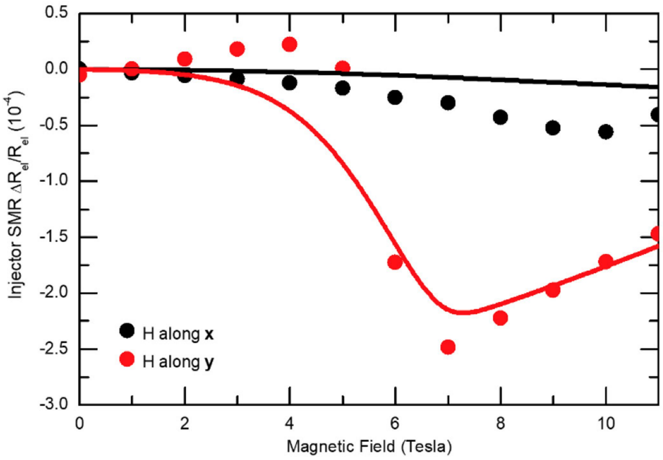
Extended Data Fig. 1 | R-plane haematite. a, Crystallographic structure of haematite α-Fe_2O_3. The single crystal that we use has an R-plane termination that makes a 33° angle with the easy axis of α-Fe_2O_3 ([001]). Above $\boldsymbol{H}_c$, the (111) plane becomes an easy plane. The green and red atoms correspond to iron and oxygen, respectively. **b**, Temperature dependence of the magnetization in the single crystal of haematite measured by SQUID. The weak ferromagnetic moment is suppressed below the Morin temperature of 260 K. Inset, at 200 K (below the Morin temperature) there is a complete absence of ferromagnetic moment.



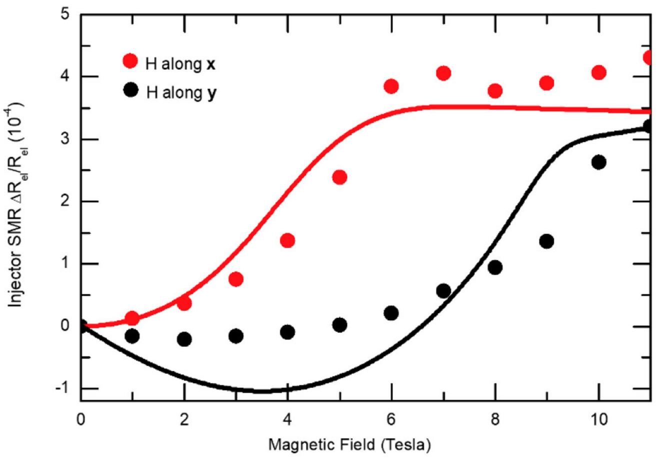
Extended Data Fig. 2 | Weak ferromagnetic (or canted antiferromagnetic) domains. Magnetic domains are imaged by MOKE microscopy above the Morin temperature (at 290 K) of an R-cut haematite single crystal.


a
z
y
x
H=0
x'
33°
n
m=0
easy axis
b
z
y
x
n⊥x'
H
m∝y
easy axis
c
z
y
x
H
①
n⊥y
m⊥y
easy axis
②
z
y
x
H
m∝x
n∝y
easy axis

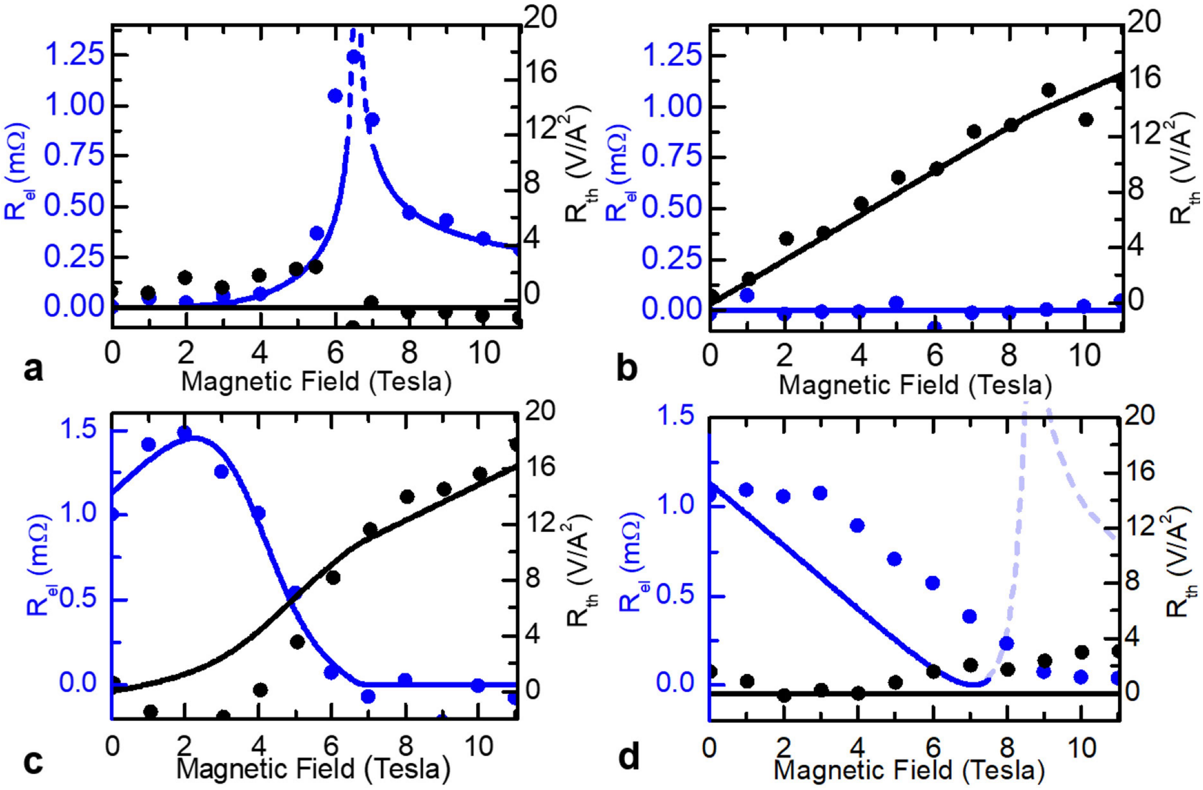
Extended Data Fig. 3 | Magnetic states of haematite under a field. **a**, Initial magnetic state of the R-cut haematite single crystal in its easy-axis antiferromagnetic phase at 200 K. The easy axis is oriented about 33° out of plane but only the in-plane projection is relevant. The projection of the easy axis in shown in (*x*, *y*, *z*) coordinates. **b**, Evolution of the antiferromagnetic state when a field is applied along ***y***. **c**, Evolution of the antiferromagnetic state when the field is applied along ***x*** at (1) small and (2) large fields ***H***.



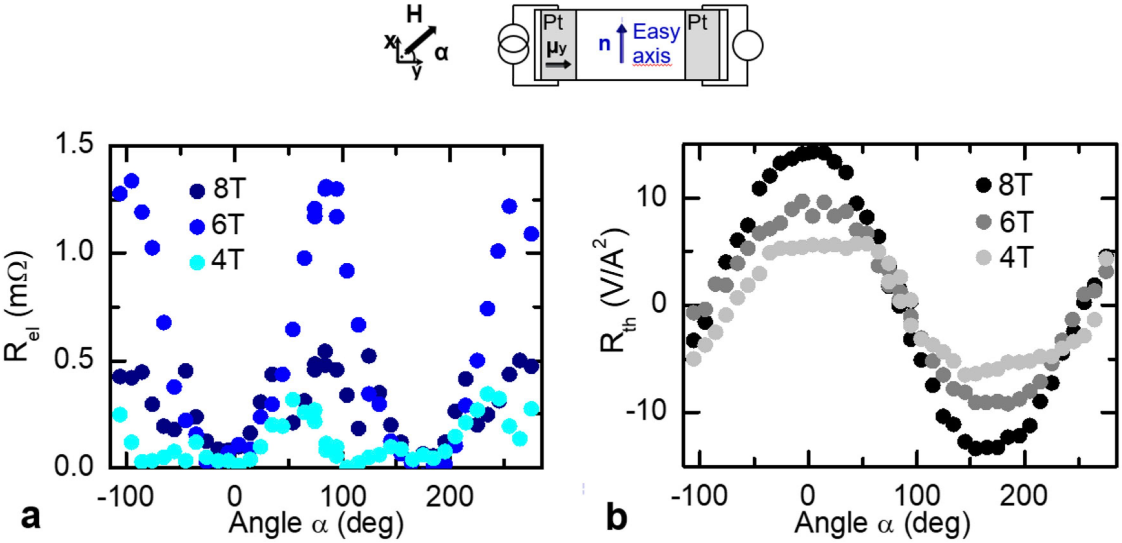
Extended Data Fig. 4 | SMR for a geometry with the wires along *x* (*μ* along *y*). The black and red circles show data for ***H*** along ***x*** and ***y***, respectively. The drop in resistance in the ***H***-along-***y*** data at about 6 T corresponds to $\boldsymbol{H}_{\mathrm{c}}$. The solid lines correspond to a combined regression fit of the ***H***-along-***x*** and ***H***-along-***y*** data of the spin-mixing conductances from equations (4)–(6), yielding $g_{\boldsymbol{n}}^{(\uparrow\downarrow)} = 12.76\ \mathrm{nm}^{-2}$ and $g_{\mathrm{r}\boldsymbol{m}}^{(\uparrow\downarrow)} = 74.82\ \mathrm{nm}^{-2}$. The two field directions are fitted together with $g_{\mathrm{i}\boldsymbol{m}}^{(\uparrow\downarrow)} = 0$, $K/J = 0.025$ and $D/J = 0.011$, using $\lambda = 2$ nm, $\theta_{\mathrm{SH}} = 0.1$ and $\rho_0 = 40\ \mu\Omega$ cm for the normal metal wires.



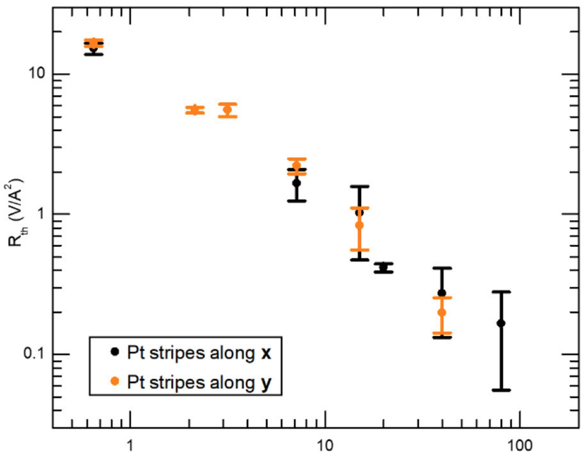
Extended Data Fig. 5 | SMR for a geometry with the wires along *y* (μ along *x*). The red and black circles show data for ***H*** along ***x*** (perpendicular to the wire) and along ***y*** (parallel to the wire), respectively. The jump in resistance around 8–10 T for ***H*** parallel to the wire (black) corresponds to the DMI-induced $\boldsymbol{H}_c$ (Methods section ‘Modelling of the Néel order and magnetization orientations’). The jump in resistance at 6 T for ***H*** perpendicular to the wire (red) correspond to the more classical $\boldsymbol{H}_c$. The solid lines correspond to a combined regression fit of the ***H***-along-***x*** and ***H***-along-***y*** data of the spin-mixing conductances from equations (4)–(6), yielding $g_{\boldsymbol{n}}^{(\uparrow\downarrow)} = 31.77\ \text{nm}^{-2}$ and $g_{\text{r}\boldsymbol{m}}^{(\uparrow\downarrow)} = 136.52\ \text{nm}^{-2}$. The two field directions are fitted together with $g_{\text{i}\boldsymbol{m}}^{(\uparrow\downarrow)} = 0$, $K/J = 0.025$ and $D/J = 0.011$, using $\lambda = 2$ nm, $\theta_{\text{SH}} = 0.1$ and $\rho_0 = 40\ \mu\Omega$ cm for the normal metal wires.



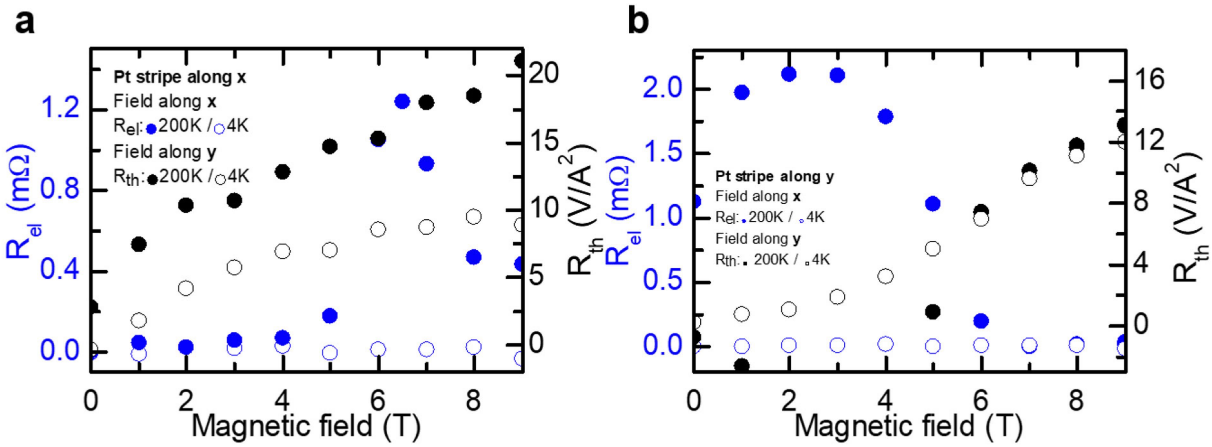
Extended Data Fig. 6 | Fitting of spin-transport signals. For the fits we use equation (6) and $S_{\boldsymbol{m}} \approx S_0$, with $G_{\boldsymbol{m}} = S_{\boldsymbol{n}} = 0$, $K/J = 0.025$ and $D/J = 0.011$. **a**, **b**, Platinum wires along $\boldsymbol{x}$. The fits correspond to Fig. 2. From fitting R_{el} in **a**, we obtain a value of $G_0 \approx 0.17$ mΩ. Here, fitting is done by excluding the largest three points of R_{el} (dotted lines), where the stochastic transport theory[12] is expected to break down owing to strong magnon scattering processes. The main peak around the $\boldsymbol{H}_{\mathrm{c}}$ corresponds to a reduction in the gap Δ_{c} of one magnon mode. In fitting **b**, we extract $S_0 \approx 60.6$ V A^{-2} from R_{th} induced by $\boldsymbol{m}$. **c**, **d**, Platinum wires along $\boldsymbol{y}$. The fits correspond to Fig. 3c. From fitting R_{th} in **c**, we extract $S_0 \approx 64.3$ V A^{-2}, in agreement with the value from **b**. From fitting R_{el} in **c**, we obtain $G_0 \approx 1.42$ mΩ (contrary to the value for **a**). One explanation for this discrepancy could be a different spin-mixing conductance for this orientation, which would be consistent with the spin-mixing conductances extracted from the SMR data (Extended Data Figs. 4, 5). Our theoretical model in **d** introduces features not observed in the high-field experimental data; hence, we fit the data using $G_0 \approx 1.42$ mΩ from **c**. The pale blue dashed line indicates the theoretical features that were not observed in the experiment.



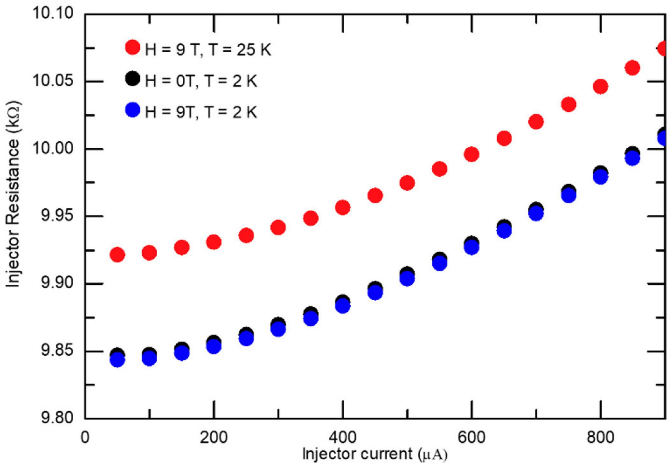
Extended Data Fig. 7 | Angular dependence of the non-local voltages for wires along *x* at different *H* (below, at and above H_c). **a**, R_{el} has a $\sin^2(\alpha)$ angular dependence, as expected for a spin signal. **b**, R_{th} has an $\sin(\alpha)$ angular dependence, as expected for a thermally injected signal. The two signals have maxima shifted by 90°, which shows that R_{el} corresponds to magnons injected along ***n*** whereas R_{th} corresponds to magnons injected along ***m***.



Extended Data Fig. 8 | Distance dependence of R_{th} for H perpendicular to the wires (along either x or y). At a field of 10 T, far above H_c, R_{th} does not depend on the orientation of the wire. Error bars are a combination of the standard error on the mean and the measurement noise floor of the system.



Extended Data Fig. 9 | Spin-transport at 4 K and 200 K. a, **b**, Platinum wires along ***x*** and ***y***, respectively. R_{el} is shown for ***H*** applied along ***x*** (the easy axis) at 200 K (filled blue circles) and 4 K (open blue circles). R_{th} is shown for ***H*** applied along ***y*** at 200 K (filled black circles) and 4 K (open black circles). The distance between wires is about 200 nm (**a**) or 500 nm (**b**), the wire lengths are 80 μm (**a**) or 160 μm (**b**) and the applied current is 300 μA (**a**) or 600 μA (**b**).



Extended Data Fig. 10 | Resistance of the platinum injector as a function of injection current and temperature. The resistance of platinum is a reliable indicator of the local temperature of the area. An injection current of 600 μA (current density of 2.8×10^{11} A m^{-2}) at 2 K corresponds to the same local temperature as the injector in a helium bath of 25 K with an injection current of 50 μA (current density of 2.3×10^{10} A m^{-2}), signified by the identical resistance values at these two points and hence the same local temperature.

LETTER

https://doi.org/10.1038/s41586-018-0474-7

Three-dimensional printing of hierarchical liquid-crystal-polymer structures

Silvan Gantenbein[1], Kunal Masania[1]*, Wilhelm Woigk[1], Jens P. W. Sesseg[2], Theo A. Tervoort[2]* & André R. Studart[1]*

Fibre-reinforced polymer structures are often used when stiff lightweight materials are required, such as in aircraft, vehicles and biomedical implants. Despite their very high stiffness and strength[1], such lightweight materials require energy- and labour-intensive fabrication processes[2], exhibit typically brittle fracture and are difficult to shape and recycle[3,4]. This is in stark contrast to lightweight biological materials such as bone, silk and wood, which form by directed self-assembly into complex, hierarchically structured shapes with outstanding mechanical properties[5–11], and are circularly integrated into the environment. Here we demonstrate a three-dimensional (3D) printing approach to generate recyclable lightweight structures with hierarchical architectures, complex geometries and unprecedented stiffness and toughness. Their features arise from the self-assembly of liquid-crystal polymer molecules into highly oriented domains during extrusion of the molten feedstock material. By orienting the molecular domains with the print path, we are able to reinforce the polymer structure according to the expected mechanical stresses, leading to stiffness, strength and toughness that outperform state-of-the-art 3D-printed polymers by an order of magnitude and are comparable with the highest-performance lightweight composites[1,12]. The ability to combine the top-down shaping freedom of 3D printing with bottom-up molecular control over polymer orientation opens up the possibility to freely design and realize structures without the typical restrictions of current manufacturing processes.

The manufacture of polymer-based lightweight structures is currently limited either to 3D-printed complex parts, which show poor molecular orientation and mechanical strength[13,14] (Extended Data Fig. 1), or to highly oriented stiff polymers, the shapes of which are restricted to simple geometries[15,16]. To combine shaping freedom with molecular orientation, the 3D printing of liquid-crystal elastomers has been recently exploited[17–19]. Although desirable shape-morphing effects can be achieved, the Young's modulus of such soft elastomers is three to four orders of magnitude lower than those of high-performance liquid-crystal synthetic fibres. This is explained by the presence of relatively long, flexible segments in the molecular structure of the polymer.

To fully exploit the shaping freedom of 3D printing and the favourable mechanical properties expected from molecularly oriented liquid-crystal polymers (LCPs), we propose an approach that conceptually follows two design principles that are used in nature to grow strong and tough biological materials[20]. First, anisotropy is achieved during the printing process by self-assembly of the LCP along the print path. Second, anisotropy and the complex-shaping capabilities offered by 3D printing are exploited to tailor the local stiffness and strength of the structure according to the specific loading conditions imposed by the environment.

The rigid molecular segments of aromatic thermotropic polyesters are known to self-assemble into nematic domains at temperatures higher than the melting temperature of the material (Fig. 1a, b). In the polymer melt at rest, there is no preferred orientation of the individual ordered nematic domains, leading to a globally random orientation distribution (Fig. 1c). Extrusion of the polymer melt through the nozzle of the 3D printer gives rise to shear and extensional flow fields that align the nematic domains in the direction of flow (Fig. 1d). As the molten polymer exits the nozzle, the flow ceases and the extruded filament is exposed to the ambient temperature. A temperature gradient is then formed between the cold surface of the filament and its hot interior. Rapid cooling at the surface leads to solidification of the nematic order in the flow-aligned arrangement. By contrast, polymer chains present in the interior of the filament experience slower cooling, and therefore have sufficient time to re-orient, driven by thermal motion. As a result, the extruded filaments possess a core–shell structure, in which a highly aligned skin encloses a less oriented core (Figs. 1e, f, 2a). The thickness of the skin depends on the filament diameter and the operating temperatures. A simple analytical heat-transfer model was used to predict the effect of these relevant printing parameters on the final core–shell architecture (Fig. 2b and equation (16) in Methods). Optical microscopy and X-ray scattering experiments confirm the formation of a highly aligned skin, the relative fraction of which increases as the filament size decreases (Fig. 2b–d, Extended Data Fig. 2). Although the possibility of aligning thermotropic LCPs in the nozzle of a fused deposition modelling (FDM) printer has been suggested previously, the mechanical properties of the obtained objects remained below those of their conventional injection-moulded counterparts[21]. By contrast, the core–shell filaments produced in our study showed notable mechanical strength and elastic modulus (Fig. 2).

The effect of the core–shell architecture on the tensile mechanical behaviour of individual filaments was investigated in order to establish a quantitative correlation between mechanical performance and relevant printing parameters. Individual filaments can be extruded vertically starting from the build platform, or horizontally onto the substrate at a constant height. The vertically extruded filament cools symmetrically in air and its diameter is controlled by the diameter of the nozzle (Fig. 1e). During the deposition of horizontal filaments, inhomogeneous heat-transfer conditions at the boundary of the filament and the substrate result in non-uniform skin thickness in the filament cross-section (Fig. 1f). Regardless of the printing configuration, the Young's modulus of the printed filaments increases as much as threefold when the nozzle diameter is decreased from 800 μm to 150 μm, and as much as sevenfold when the layer height is decreased from 400 μm to 50 μm (Fig. 2e, f). A similar trend was observed for the tensile strength of the filaments. This can be qualitatively explained by the increase in the relative fraction of aligned molecules as the filament diameter or layer height decreases. For vertical filaments, the dependence of the Young's modulus on the nozzle diameter is well described by a simple scaling law that is derived from the heat-transfer model used to predict the thickness of the skin layer (equation (16) in Methods). According to this model, the Young's modulus (E) scales with the inverse of the nozzle diameter ($1/d_N$). The ability to produce filaments thinner than the nozzle diameter and thus increase considerably the fraction of aligned domains (Fig. 2c) makes the horizontal

---

[1]Complex Materials, Department of Materials, ETH Zürich, Zürich, Switzerland. [2]Soft Materials, Department of Materials, ETH Zürich, Zürich, Switzerland. *e-mail: kunal.masania@mat.ethz.ch; theo.tervoort@mat.ethz.ch; andre.studart@mat.ethz.ch

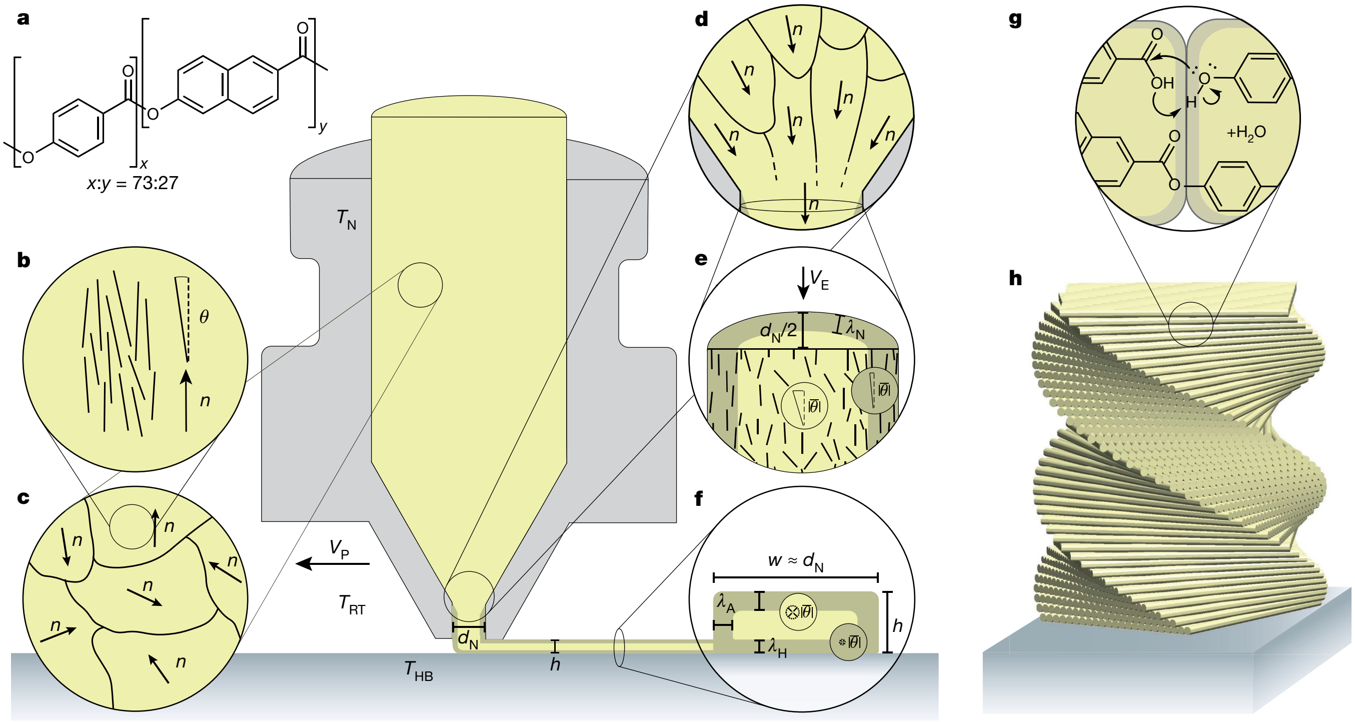


Fig. 1 | Printing hierarchical, thermotropic LCPs using fused deposition modelling. a, An aromatic random copolyester consisting of the rigid monomers *p*-hydroxybenzoic acid and 2-hydroxy-6-naphthoic acid forms short rod-like polymer chains. **b**, The rigid polymer rods align along the same director ***n*** in the melt. They are not perfectly oriented but show a certain misalignment θ towards the director, forming a nematically ordered domain. **c**, Locally aligned nematic domains with randomly oriented directors form the quasi-isotropic bulk of the polymer. **d**, The directors are subjected to elongational and shear forces during the extrusion through the heated nozzle (at temperature T_N) with diameter d_N. These forces can rearrange the polymers in the melt and align the directors along the extrusion direction. **e**, The material starts to lose its orientation once it has been extruded. At the same time the solidification front freezes the nematic order in place, starting from the surface. These two effects result in a core–shell microstructure with a highly aligned shell of thickness λ_N. **f**, Depositing material on a surface at a height *h* promotes further alignment of the director in the print direction and forms a core–shell structure with different skin thicknesses due to the distinct cooling rates for air, the glass surface and the polymer substrate. T_{HB}, temperature of the heated bed or substrate; T_{RT}, temperature of the environment (room temperature); V_E, velocity of extrusion; V_P printing velocity; *w*, print width; λ_A, thickness of skin aligned towards air; λ_H, thickness of skin aligned towards the heated bed or substrate. **g**, **h**, After printing, chain ends can chemically crosslink via thermal annealing, which increases the molecular weight and enhances stress transfer between filaments.

printing configuration very effective for improving the stiffness and the strength of the printed materials.

In addition to the nozzle diameter and layer height, temperature is another parameter that can be controlled conveniently during the printing process. Our experimental data show that both the Young's modulus and the tensile strength of printed filaments increase as the temperature at the nozzle is decreased. This correlation can be qualitatively explained by the fact that the solidification process is faster if the nozzle temperature is closer to the melting temperature of the polymer. Faster solidification results in thicker skins and therefore improved stiffness and strength. This effect can also be quantitatively illustrated using the scaling law obtained from the analytical heat-transfer model (Fig. 2g and equation (16) in Methods).

An important additional feature of the aromatic polyester used as the liquid-crystal feedstock is that its printed filaments can be chemically cross-linked by thermal annealing. Cross-linking occurs via post-condensation reactions between carboxylic acid and hydroxy groups present at the end of the macromolecules (Fig. 1g). The formation of cross-links is expected to increase the molecular weight and enhance stress transfer between adjacent macromolecules, thereby improving the tensile behaviour of the polymer under load. Indeed, annealing for 96 h was found to increase the tensile strength of printed filaments by a factor of two, reaching a value of 1 GPa for filaments printed horizontally. Within the first 10 h of thermal treatment, the tensile strength of the filament was observed to follow a square-root dependence on the annealing time. This suggests that the cross-linking reaction is controlled by the diffusion of reaction products—such as water molecules—away from the reaction sites, which supports the hypothesis that a post-condensation reaction is responsible for the observed strengthening effect (Extended Data Fig. 3a).

After analysing the structure and properties of individual filaments, we then turned our attention to more complex fibre laminates that are obtained by printing filaments adjacent to one another in a specific direction (Fig. 3). As expected, maximal Young's modulus and tensile strength are achieved by printing filaments parallel to the mechanical-loading direction (Fig. 3a). Increasing the angle between filament orientation and the loading direction decreases the Young's modulus of the printed structure substantially, as predicted from classical laminate theory (Fig. 3a, Extended Data Fig. 4 and equation (18) in Methods). Overall, the structure, stiffness and strength of the printed laminates can be tuned over a wide range using the same control parameters that were investigated for individual filaments, namely layer height and temperature (Extended Data Fig. 5).

The molecular cross-links formed by thermal annealing also have a considerable effect on the strength of the printed laminates. For low alignment angles (0° and 30°), the bending strength of the laminate increases by 50% to 100% after the thermal treatment (Fig. 3b). Annealed samples tested in tensile mode at 90° also show an almost 100% increase in strength after 96 h (Fig. 3c), while simultaneously increasing the Young's modulus owing to enhanced load transfer between filaments in the composite. The observed strengthening effect follows the diffusion-controlled kinetics that are expected for the post-condensation reaction (Extended Data Fig. 3b). More importantly, thermally induced cross-linking markedly changes the fracture

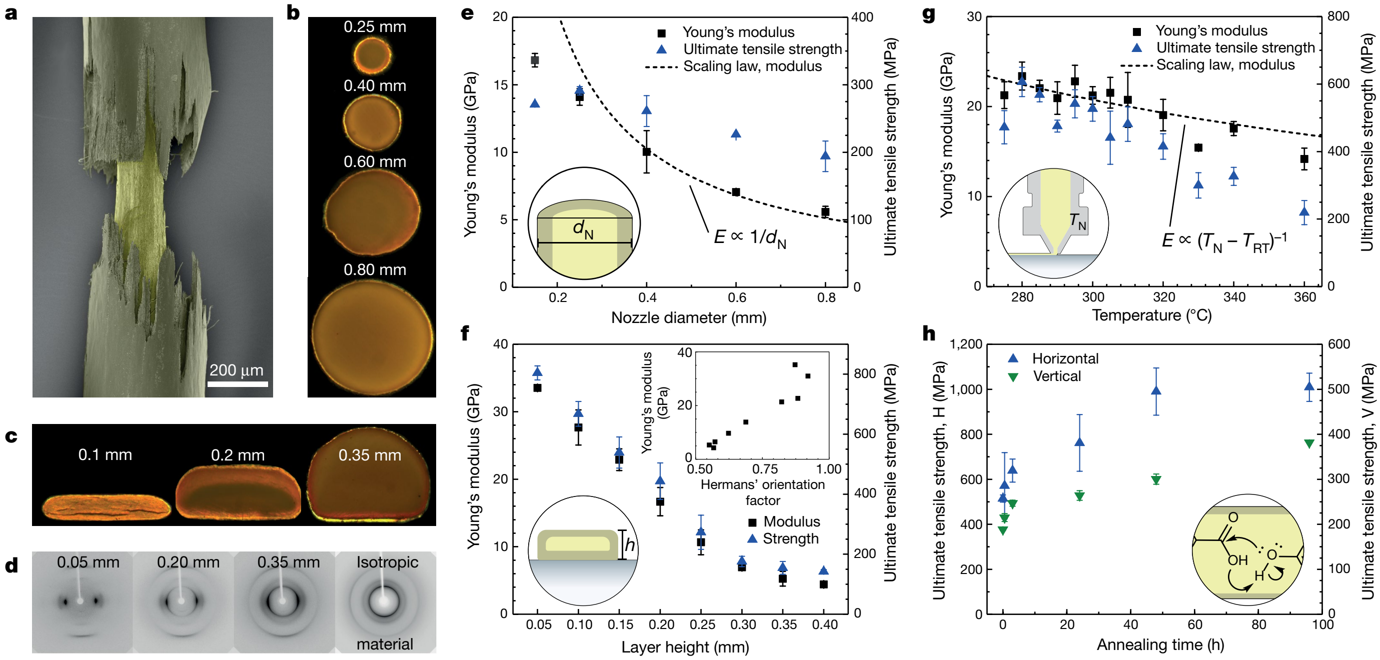


Fig. 2 | The LCP filament properties correlate strongly with the printing conditions. a, False-coloured scanning electron microscopy image of the tensile-tested vertical filament confirms a core–shell structure. The fact that the core of the fibre is still intact indicates that the shell of the LCP fibre is the stiffer phase. **b**, **c**, Polarized light microscopy of a 100-μm cross-section confirms the presence of a core–shell structure in both vertical (**b**) and horizontal (**c**) filaments, as indicated by the more intense illumination of the skin of the printed filaments compared to the core. **d**, X-ray diffraction measurements confirm a higher fraction of oriented polymers in thinner samples. **e**, The Young's modulus and the strength of vertically extruded lines increase with decreasing nozzle diameter because of the increased fraction of oriented polymer in the print line. This hypothesis is supported by a scaling law that successfully captures the dependence of the Young's modulus on the nozzle diameter d_N. **f**, This effect of the core–shell architecture also translates to horizontally printed filaments, where the Young's modulus and strength increase with decreasing print filament height, reaching a maximum reported modulus of 34 GPa and a maximum strength of 800 MPa. **g**, Increased printing temperatures lead to longer cooling times, and therefore longer times for the freezing of the aligned nematic microstructure. This results in an inverse correlation between the Young's modulus and the temperature difference ($T_N - T_{RT}$). **h**, Solid-state annealing further enhances the properties by increasing the molecular weight of the polymer. The strength increases up to 400 MPa for vertically extruded filaments (V) and up to 1 GPa for horizontally printed filaments (H). Error bars indicate the standard deviation for five to ten measurements at each data point.

behaviour of the printed laminates (Fig. 3d). Complete fracture of laminates subjected to annealing is possible only after more than 8% tensile strain is applied. During such straining, multiple peak loads have to be overcome, which leads to substantial energy dissipation during the fracture process. The multiple stress peaks observed resemble those of toughening mechanisms in resilient biological materials such as mollusc shells and bone. This toughening effect contrasts with the abrupt failure of the unannealed specimens after a peak stress is achieved. On the basis of these results and the increased shear strength of the laminates (Extended Data Fig. 6a), we infer that the thermal treatment not only strengthens the individual filaments but also establishes cross-links between filaments in the laminate. Such cross-links are expected to enhance stress transfer between filaments (Extended Data Fig. 6b) and to prevent their delamination through crack-arresting mechanisms. Therefore, the high toughness of the annealed laminates probably emerges from the hierarchical cross-linking of macromolecules and of filaments at increasing length scales.

Besides the self-assembly and hierarchical cross-linking of macromolecules, our 3D printing approach provides a means to replicate other design principles of biological materials. Analogous to the biological growth of living tissues, 3D printing is unique in its ability to create customized material architectures in a layer-by-layer additive fashion. We illustrate this capability by 3D-printing laminate architectures in which the filament orientation is designed to best respond to the local strains imposed by a specific mechanical loading condition. In our example, a plate with a central hole is subjected to a global tensile strain in one specific direction (Fig. 4a). Two distinct laminate architectures containing either straight (unidirectional) or curvilinear (directional) filaments (Extended Data Fig. 7) were tested and compared to a reference sample that comprised isotropic hot-pressed LCPs.

The higher intrinsic strength and stiffness of printed filaments leads to a threefold increase in stiffness and a fivefold increase in strength of the unidirectional architecture compared to isotropic LCPs. The use of a load-specific print architecture that follows the stress lines that develop around the hole doubles the strength of the sample compared to the unidirectional sample. Thermal annealing further increases the strength and stiffness of both sample geometries.

Laminates with unidirectional and curvilinear designs contain filaments lying predominantly parallel to the loading direction to arrest cracks that propagate from the central hole, which acts as a stress concentrator. The high local strains in the periphery of the hole were experimentally confirmed by strain maps obtained by digital image correlation analysis (Fig. 4b). Similar to the fibre architectures found in wood and bone, the aligned reinforcement of such laminates prevents catastrophic failure by promoting crack-arresting mechanisms within the material as well as providing strength and toughness via the core–shell structure of the printed filament. As a result, the modulus of toughness of laminates with aligned architectures is about 30- to 55-fold higher than that of the isotropic reference specimen. The higher strength and Young's modulus of specimens that display the curvilinear reinforcement design demonstrates our ability to produce high-performance laminates by tuning the fibre orientation to best match the stress lines developed throughout the mechanically loaded structure.

The high toughness of the printed laminates is accompanied by a marked increase in specific Young's modulus without loss of the intrinsically high damping properties of aromatic polyesters (Fig. 4c). With this unique combination of properties, our printed LCP laminates clearly stand out in comparison to other existing lightweight materials. In terms of specific strength and Young's modulus, printed LCPs surpass the properties of glass fibre-reinforced polymers and

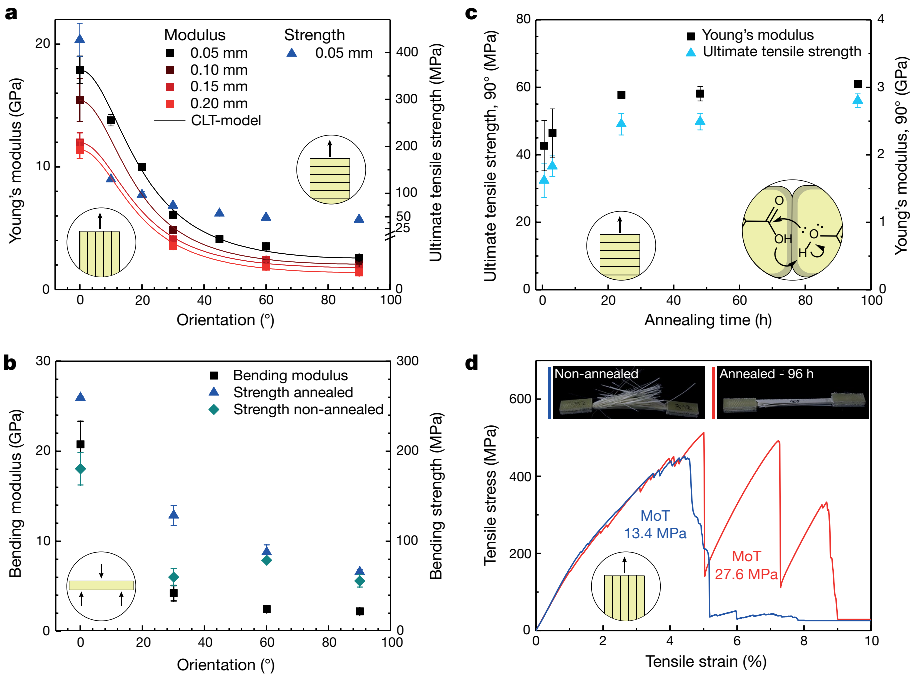


Fig. 3 | Mechanical properties of LCP parts printed through the unidirectional deposition of filaments. a, Like that of fibre-reinforced composite materials, the Young's modulus of unidirectional printed parts depends on the printing orientation, as predicted by classical laminate theory (CLT). **b**, Bending properties follow the same trend, as expected for a system composed of oriented fibres. Annealing increases the bending strength by increasing the stress transfer between different print filaments. **c**, Annealing tensile samples printed 90° to the loading direction increases the strength over time owing to the improved adhesion between filaments. Initially, this process is controlled by the diffusion of water molecules generated during the post-condensation reaction. Error bars indicate the standard deviation for five to ten measurements at each data point. **d**, For samples with print lines oriented in the loading direction, the fracture mode changes from a broom-like to a tough layer-like fracture owing to the improved filament adhesion achieved by thermal annealing (96 h). This change also results in a saw-tooth break pattern in the stress–strain curve, which increases the amount of energy required to fracture the part (modulus of toughness, MoT) by a factor of two.

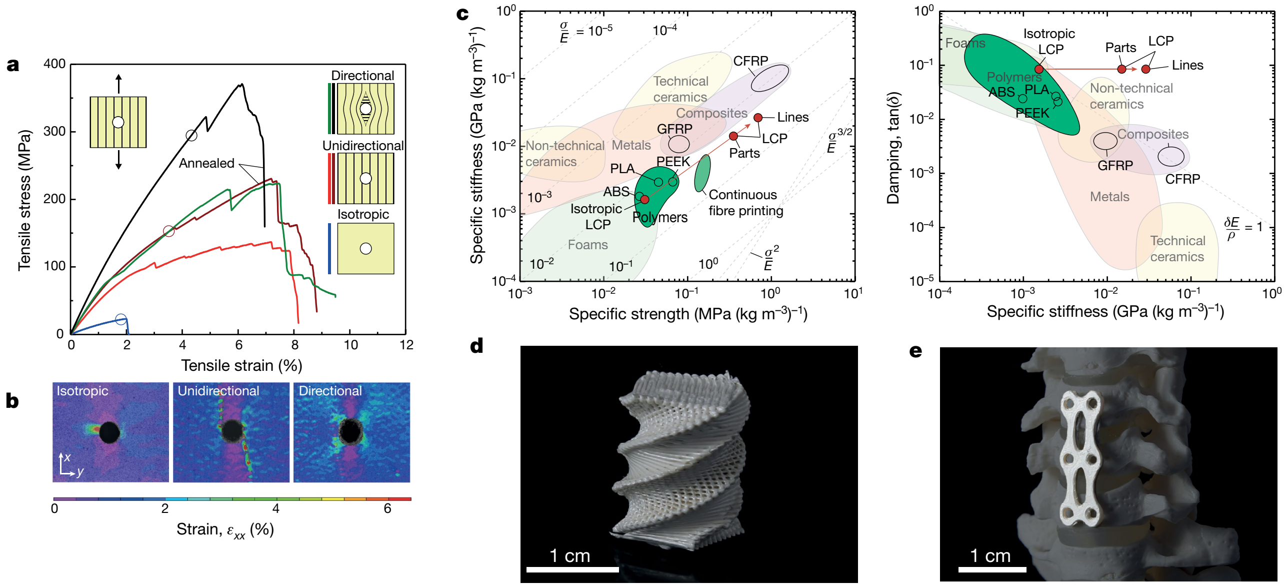


Fig. 4 | Mechanical properties and complex geometry of 3D-printed LCP laminates and parts. a, Mechanical response under tension of open-hole LCP laminates, highlighting the improved properties obtained when the filament architecture is designed to follow the stress lines developed within the loaded material. **b**, Open-hole strain maps measured by digital image correlation just before fracture. **c**, Ashby diagrams indicating the specific stiffness and specific strength (left) and damping (right) properties of LCP lines and parts compared to the isotropic counterpart and state-of-the-art polymers and reinforced composite materials[22]. σ, strength; ρ, density; ABS, acrylonitrile butadiene styrene; CFRP, carbon fibre reinforced polymer; GFRP, glass fibre reinforced polymer; PEEK, polyether ether ketone; PLA, polylactic acid. **d**, **e**, Examples of 3D-printed LCP parts with complex fibre architectures and geometries: impact-resistant Bouligand-type structure with twisted plywood arrangement of printed fibres (**d**) and biomedical implant with a local bearing enhancement where the print lines are programmed to follow the principle stress direction around holes (**e**).

continuous fibre-printed composites (Fig. 4c). Indeed, the stiffness and strength of our printed laminates approach those of carbon-fibre reinforced polymers, with the added value of recyclability (Extended Data Fig. 8), automated manufacturing and lower carbon footprint. By taking advantage of the shaping capabilities of 3D printing, special complex geometries and application-specific shapes can be produced using the proposed additive technology (Fig. 4d, e). By combining the tool path control provided by the 3D printer with the tunable orientation of self-assembled building blocks within the ink, we can reach unparalleled levels of hierarchical structural complexity (Fig. 4d). As opposed to current FDM printed polymer parts, the anisotropic filament architectures implemented in these complex geometries can be designed to achieve toughness comparable to that of the strongest polymer-based lightweight materials. This approach is in contrast with the usual strategy to reduce anisotropy in 3D-printed synthetic parts, and opens up the possibility to fabricate structures that can fulfil different performance requirements while ensuring a sustainable circular life cycle of the material.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0474-7.

Received: 12 March 2018; Accepted: 19 July 2018;
Published online 12 September 2018.

1. Amacher, R. et al. Thin ply composites: experimental characterization and modeling of size-effects. *Compos. Sci. Technol.* **101**, 121–132 (2014).
2. Witik, R. A., Payet, J., Michaud, V., Ludwig, C. & Månson, J.-A. E. Assessing the life cycle costs and environmental performance of lightweight materials in automobile applications. *Composites, Part A* **42**, 1694–1709 (2011).
3. Kiser, B. Circular economy: Getting the circulation going. *Nature* **531**, 443–446 (2016).
4. Yang, Y. et al. Recycling of composite materials. *Chem. Eng. Process.* **51**, 53–68 (2012).
5. Launey, M. E., Buehler, M. J. & Ritchie, R. O. On the mechanistic origins of toughness in bone. *Ann. Rev. Mater. Res.* **40**, 25–53 (2010).
6. Wegst, U. G. K., Bai, H., Saiz, E., Tomsia, A. P. & Ritchie, R. O. Bioinspired structural materials. *Nat. Mater.* **14**, 23–36 (2015).
7. Meyers, M. A., Chen, P. Y., Lin, A. Y. M. & Seki, Y. Biological materials: Structure and mechanical properties. *Prog. Mater. Sci.* **53**, 1–206 (2008).
8. Burgert, I. Exploring the micromechanical design of plant cell walls. *Am. J. Bot.* **93**, 1391–1401 (2006).
9. Cranford, S. W., Tarakanova, A., Pugno, N. M. & Buehler, M. J. Nonlinear material behaviour of spider silk yields robust webs. *Nature* **482**, 72–76 (2012).
10. Vollrath, F. & Knight, D. P. Liquid crystalline spinning of spider silk. *Nature* **410**, 541–548 (2001).
11. Smith, B. L. et al. Molecular mechanistic origin of the toughness of natural adhesives, fibres and composites. *Nature* **399**, 761–763 (1999).
12. Grossman, M. et al. Mineral nano-interconnectivity stiffens and toughens nacre-like composite materials. *Adv. Mater.* **29**, 1605039 (2017).
13. Ahn, S. H., Montero, M., Odell, D., Roundy, S. & Wright, P. K. Anisotropic material properties of fused deposition modeling ABS. *Rapid Prototyping J.* **8**, 248–257 (2002).
14. Ning, F., Cong, W., Qiu, J., Wei, J. & Wang, S. Additive manufacturing of carbon fiber reinforced thermoplastic composites using fused deposition modeling. *Composites, Part B* **80**, 369–378 (2015).
15. Lefèvre, J. et al. "Foil spintrusion" of high-performance polymer films. *J. Polym. Sci. B* **50**, 1713–1727 (2012).
16. Schaller, R., Peijs, T. & Tervoort, T. A. High-performance liquid-crystalline polymer films for monolithic "composites". *Composites, Part A* **81**, 296–304 (2016).
17. Kotikian, A., Truby, R. L., Boley, J. W., White, T. J. & Lewis, J. A. 3D printing of liquid crystal elastomeric actuators with spatially programed nematic order. *Adv. Mater.* **30**, 1–6 (2018).
18. Ambulo, C. P. et al. Four-dimensional printing of liquid crystal elastomers. *ACS Appl. Mater. Interfaces* **9**, 37332–37339 (2017).
19. López-Valdeolivas, M., Liu, D., Broer, D. J. & Sánchez-Somolinos, C. 4D printed actuators with soft-robotic functions. *Macromol. Rapid Commun.* **39**, 1700710 (2018).
20. Fratzl, P. & Weinkamer, R. Nature's hierarchical materials. *Prog. Mater. Sci.* **52**, 1263–1334 (2007).
21. Gray, R. W., Baird, D. G. & Bøhn, J. H. Rapid prototyping journal effects of processing conditions on short TLCP fiber reinforced FDM parts. *Rapid Prototyping J.* **4**, 14–25 (1998).
22. Ashby, M. F. *Material and Process Selection Charts* http://www.grantadesign.com/download/pdf/teaching_resource_books/2-Materials-Charts-2010.pdf (Granta Design, 2010).

Acknowledgements We thank G. Ghazaryan and the Kunststoff Ausbildungs- und Technologie-Zentrum for support with filament extrusion, the D-MATL X-ray Platform for access to the diffractometer, and J. Vermant, K. Feldman and N. Bahamonde for discussions. The Swiss Competence Center for Energy Research (SCCER - Capacity Area A3: Minimization of Energy Demand), ETH Foundation grant SP-MaP 01-15, SNSF Project 200021_156011 and consolidator grant BSCGIO_157696 are acknowledged for supporting this research.

Author contributions T.A.T. and A.R.S. conceived the idea together with S.G. and K.M.; S.G., W.W., K.M. and J.S. carried out the experimental work; S.G., K.M., W.W., T.A.T. and A.R.S. carried out the analysis and co-wrote the paper.

Competing interests The authors have filed patent application EP18179376 relating to this work.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0474-7.

Reprints and permissions information is available at http://www.nature.com/reprints.

Correspondence and requests for materials should be addressed to K.M., T.A.T. or A.R.S.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Preparation of LCP filament. LCP (Vectra A950, Ticona) was dried at 150 °C for 12 h before extrusion. LCP filaments were prepared using a single-screw extruder (Teach-Line E20 T, Collin) that was heated to 280, 290, 290 and 260 °C at the four zones along its longitudinal axis. The filament was extruded at 60 r.p.m., cooled in a water bath (Teach-Line WB850, Collin) and collected on a modified flat film line (Teach-Line CR72 T, Collin). The speed of the collector rolls was adjusted to achieve a filament diameter of 1.75 mm. The filament was wound on FDM spools and dried at 70 °C for at least 24 h before use. Because flow-induced molecular orientation is a well-established phenomenon for liquid-crystal systems[23–27], the proposed printing technology should be applicable to a broad range of LCPs beyond the polyester used in this study.

FDM printing setup. A commercially available FDM printer (Ultimaker 2+, Ultimaker) was modified with a geared direct drive extruder and an all-metal V6 hotend (E3D) to achieve temperatures up to 400 °C (Extended Data Fig. 9). The borosilicate glass build-plate was heated to 90 °C and coated with a thin layer of PVA-based adhesive spray (3DLac) before printing with LCP to improve bed adhesion and reduce warping. Generally, parts were printed at 295 °C at a speed of 35 mm s^{-1} with part cooling fans running at 20%. For printing lines, the speed was reduced to 20 mm s^{-1} to improve the quality of lines in contact with the glass surface.

PLA and PEEK reference samples were printed using commercial filaments (Dutch Filaments B.V. and 3D4MAKERS) at a temperature of 210 °C and 380 °C, respectively. The build-plate was heated to 60 °C for PLA and 120 °C for PEEK. All other parameters were kept the same as for printing LCP. PEEK samples were further heat-treated after printing at 150 °C for one hour followed by one hour at 200 °C to reach optimum crystallization.

Print paths (Gcode) with reduced control over the print directions were generated with Cura, an open-source FDM slicer from Ultimaker. A custom slicer using Grasshopper for Rhinoceros (McNeel) was developed for objects with spatially tuned directional print paths or those for which the orientation of the print path was important.

Thermal annealing. The solid-state thermal annealing was performed by heating the samples to 270 °C under constant nitrogen flow for 0 h to 96 h. Solid-state cross-linking occurs via post-condensation reactions between carboxylic acid and hydroxy groups[28] of the printed samples. The samples were fixed on a steel plate with polyimide tape to prevent deformation during the annealing process.

Tensile tests of printed filaments. Tensile tests were performed on filaments printed vertically starting on the surface and moving upwards (z direction) and horizontally on the glass surface. To produce vertical filaments that do not curl during free-form extrusion, the feedstock material was first printed horizontally to ensure good attachment to the substrate. Vertical and horizontal filaments were printed with varying nozzle diameter and distance to the surface, respectively. Furthermore, the effects of the nozzle temperature and annealing time were examined for both printing configurations.

The filament samples were glued onto individual paper frames to assure a constant gauge length of 20 mm. The tensile tests were carried out at a rate of 2 mm min^{-1} using an AGS-X (Shimadzu) universal testing machine with a 1-kN capacity load cell. Five to ten samples were tested for each batch. The samples were imaged with an optical stereo microscope (WILD M10, Leica) during testing and their width and thickness measured using Fiji image analysis[29]. Data analysis was performed using a custom MATLAB script.

X-ray diffraction of printed filaments. The X-ray diffraction patterns of printed filaments with different diameters (vertical) and layer heights (horizontal) were measured on a single crystal diffractometer (Xcalibur PX, Oxford Diffraction). A molybdenum (Mo-Kα) X-ray source was used with a graphite monochromator. The samples were mounted with the director axis perpendicular to the incident X-ray beam and the scattering was detected on an Onyx CCD detector. Azimuthal angle intensity distribution $I(\phi)$ scans at $2\theta \approx 9°$ were analysed using the CrysAlisPro 171.36.20 software and the Hermans' orientation factor[30] f was calculated using the following equation:

$$f = 1 - 3\langle \cos^2\phi \rangle \qquad (1)$$

$$\langle \cos^2\phi \rangle = \frac{\int I(\phi)\sin\phi\cos^2\phi \, d\phi}{\int I(\phi)\sin\phi \, d\phi} \qquad (2)$$

Optical microscopy of printed filaments. Printed filament samples were embedded perpendicular to the print or extrusion direction using optically transparent epoxy Araldite 2020 (Huntsman) and were polished with a laboratory polishing machine (labopol-21, Struers). Polishing was carried out using progressively finer grades of emery paper at intervals of 240, 800, 1,200, 2,400 and 4,000 grit followed by 1 μm and 0.25 μm diamond pastes. The samples were then mounted on standard glass slides using the optically transparent epoxy and, finally, ground and polished to a nominal thickness of 100 μm. An AXIO Scope (Zeiss) optical microscope was used to image the cross-section of the samples using dark-field cross-polarized transmitted light.

Scanning electron microscopy of printed filaments. Printed filaments were imaged after breakage using a scanning electron microscope (Leo 1530 Gemini, Zeiss) operated with an acceleration voltage of 3 kV. Before imaging, the samples were coated with a 5 nm platinum layer using a compact coating unit (CCU-010, Safematic).

Tensile tests of printed parts. Tensile tests were carried out using a Z020 (Zwick) universal testing machine with a 20 kN capacity load cell. Unidirectional tensile test specimens (ISO 527-5) with a nominal width of 5 mm, length of 110 mm and thickness of 2 mm were printed with print-filament orientation varying from 0° to 90° with respect to the testing direction. The samples were supported in the clamping region with bonded glass fibre end tabs, resulting in a gauge length of 65 mm. Specimens were tested at a displacement rate of 2 mm min^{-1} and data was acquired for a minimum of five samples. Data analysis was performed using a custom MATLAB script. In addition to the print direction, the effects of different layer heights, temperatures, and annealing time were investigated. Additional data for the Ashby diagram in Fig. 4c were taken from the literature[1,31–35] or from https://www.carbon3d.com/materials/, or as the bulk value when data was not available.

Bending tests of printed parts. Bending tests were carried out on an AGS-X (Shimadzu) universal testing machine with a three-point bending setup using a span of 24 mm (ASTM D790M). Five repeats for samples with different thicknesses (layer heights) were measured before and after annealing at a displacement rate of 2 mm min^{-1}.

Dynamic mechanical analysis. Dynamic mechanical analysis measurements were performed on a Q800 (TA Instruments) instrument using a three-point bending set up with a 20-mm span. The printed LCP samples were 2 mm thick, 5 mm wide and 30 mm long with a 0° orientation along the span. A pre-load of 0.05 N was applied before the measurement. The displacement amplitude was set to 40 μm and the frequency was held at 100 Hz. Three samples were measured at a constant temperature of 30 °C. Five samples of the isotropic and unidirectional orientation were measured. The complex modulus E^* consisting of the storage modulus E', the loss modulus E'' and loss factor, tanδ, were measured.

Open-hole tensile test. Open-hole tensile (OHT) tests were performed on a Z020 (Zwick) universal testing machine with a 20-kN-capacity load cell. The OHT specimens had a width of 35 mm and a width-to-hole diameter ratio of 6 (ASTM D5766). Tests were conducted in accordance to the standard ISO 527. Five samples of each batch were clamped in hydraulic gripers with a clamping pressure of 5 bar and were tested in tension with a controlled displacement rate of 2 mm min^{-1}.

Video gauge and digital image correlation. The Poisson's ratio of printed LCP parts was determined by calculating the negative ratio of the transverse and longitudinal strains. Both strain components were measured simultaneously on unidirectional tensile specimens using a video gauge tracking. Optical extensometers were positioned using the Vic-Gauge software (Correlated Solutions).

The shear properties were measured on uniaxial tensile specimens with a ±45° balanced and symmetric laminate (ASTM D3518) for five repeats per batch. The shear strain was calculated as the difference between the longitudinal strain and the transverse strain of the laminate. The global strains were measured optically during the test with virtual extensometers using the Vic-Gauge software. The shear stress was calculated as the ratio of the applied load and twice the cross-sectional area of the specimen, whereas the in-plane shear strength was obtained from the applied load at failure. The in-plane shear modulus of the unidirectional lamina was obtained from the initial slope of the stress–strain curve over a strain range of 0.1%–0.5%.

Digital image correlation was used to measure the local displacements of the OHT specimens. A 12 MP Prosilica camera (Allied Vision) with an 80-mm focal length lens and a 45-mm extension tube were used to acquire images with a rate of 2 Hz. The analysis of these images was performed using the digital image correlation software Vic-2D 6 (Correlated Solutions). Analyses were carried out with a subset size and step size of 29 and 7, respectively. The software generates two-dimensional displacement fields of the loaded samples in relation to the unloaded position. These fields enable one to investigate the induced in-plane strain components, that is, ε_{xx}, ε_{yy} and γ_{xy}, as a function of the global longitudinal strain. To record the deformation of the sample surface during both tests, a speckle pattern with a dot size of about 50 μm was applied with an acryl colour spray can before testing.

Analytical models for skin formation. Two possible mechanisms may explain the formation of the core–shell architecture of the printed filaments: preferential shear-induced alignment of LCP domains at the walls of the nozzle, or rapid cooling of the filament surface. These distinct scenarios are evaluated and discussed below.

Skin formation due to shear at the nozzle wall. The formation of a skin in printed filaments may result from the shear-induced alignment of LCP molecules close to the walls of the nozzle, where shear rates are maximal during extrusion. To evaluate this hypothesis, we first estimate the thickness of such a shear-induced skin layer (λ) for different nozzle diameters (d_N). Because the skin thickness directly affects the Young's modulus of the filaments, this theoretical analysis is later compared with our experimental mechanical data to test the hypothesis.

The thickness of the skin layer depends on the shear rate profile across the radius of the nozzle. For fully developed laminar flow, the shear rate is zero at the centre of the nozzle ($r=0$) and increases continuously until it reaches the maximal value at the wall ($r=R$). Because the time required for alignment increases for lower shear rates, there will be a critical radius r_c above which the shear rates will be sufficiently high ($\dot{\gamma} > \dot{\gamma}_c$) to align the particles within the residence time of the material in the nozzle. Such a critical radius defines the skin thickness as $\lambda = R - r_c$.

To establish a correlation between λ and d_N ($2R$) we compare the time required for alignment (t_a) with the residence time (t_{res}). For an anisotropic rod-like particle suspended in a fluid, the alignment time should scale with the inverse of the shear rate, such that:

$$t \approx \frac{1}{\dot{\gamma}} \tag{3}$$

As a first approximation, we assume this simple relation to also be applicable to the LCP polymers investigated here. It follows that $t_a \approx 1/\dot{\gamma}_c$. To determine $\dot{\gamma}_c$ and thus t_a, we estimate the shear rate profile along the radius of the nozzle ($\dot{\gamma}(r)$).

The shear rate profile can be calculated from the derivative of the flow velocity with respect to the nozzle radius ($v(r)$). Such flow velocity depends on the rheological properties of the fluid. For a non-Newtonian fluid, the velocity profile is given by:

$$v(r) = \frac{3n+1}{n+1}\frac{Q}{\pi R^2}\left[1 - \left(\frac{r}{R}\right)^{1+1/n}\right] \tag{4}$$

where Q is the volumetric flow rate. Here, we assume the fluid to be described by a power-law relation, such that the viscosity (η) is given by $\eta = k\dot{\gamma}^{n-1}$, where k and n are constants. By taking the derivative of the velocity profile, the shear rate at different radial positions can be directly obtained:

$$|\dot{\gamma}(r)| = \frac{3n+1}{n}\frac{Q}{\pi R^3}\left(\frac{r}{R}\right)^{1/n} \tag{5}$$

Combining equations (3) and (5) leads to the following scaling relation for the alignment time as a function of the nozzle radius:

$$t_a \approx \frac{n}{3n+1}\frac{\pi R^3}{Q}\left(\frac{r_c}{R}\right)^{-1/n} \tag{6}$$

Comparison of the alignment time with the residence time of the material inside the nozzle enables us to predict the amount of material that aligns during the extrusion process. The residence time for a nozzle of length L is given by:

$$t_{res} = \frac{\pi R^2 L}{Q} \tag{7}$$

As shear-induced alignment occurs when $t \le t_{res}$, equations (6) and (7) can be combined to obtain:

$$\lambda = R\left\{1 - cR^n\left[\frac{n}{(3n+1)L}\right]^n\right\} \tag{8}$$

where c is a constant that is independent of the nozzle radius.

For strongly shear-thinning fluids such as the LCP melts, it is reasonable to assume that $n \to 0$. This simplifies equation (8) to $\lambda = R(1-c)$. Combining this relation with the expected Young's modulus ($E_V = \lambda E_{skin}/d_N$, see main text), we finally obtain:

$$E_V = \frac{(1-c)E_{skin}}{2} \tag{9}$$

This relation shows that the elastic modulus of the filament does not dependent on the nozzle diameter under the assumptions outlined above. Because our experimental data show a clear decrease of E_V with increasing nozzle diameters (Fig. 2e), our analysis suggests that skin formation is not controlled by shear-induced alignment at the nozzle wall.

Skin formation due to rapid cooling of the filament surface. Another hypothesis is that the thickness of the filament skin is determined by the cooling kinetics of the extrudate as the material exits the nozzle. This scenario can be theoretically described using a simple heat-transfer model. Such analysis assumes that the LCP nematic domains align strongly during extrusion of the feedstock material through the nozzle and that such alignment is maintained in the filament after extrusion only if the material is cooled fast enough to prevent random reorientation of domains driven by thermal motion. Moreover, the LCP domains in the extruded filament are assumed to reach the same level of orientation within the entire range of extrusion temperatures and filament thicknesses investigated in our experiments (Fig. 2).

The cooling of the filament after it leaves the printing nozzle can be described using the following equation describing the spatio-temporal evolution of the temperature of the material:

$$T(x,t) = T_{RT} + (T_N - T_{RT})\,\mathrm{erf}\left(\frac{x}{2\sqrt{\alpha t}}\right) \tag{10}$$

where x is the distance from the filament surface, t is the time elapsed after the material exits the nozzle, T_N is the temperature inside the nozzle, T_{RT} is the temperature of the substrate (that is, room temperature) and α is the thermal diffusivity. Using a Taylor series expansion of the error function and considering only the first term, we obtain:

$$T(x,t) = T_{RT} + (T_N - T_{RT})\left(\frac{x}{\sqrt{\pi\alpha t}}\right) \tag{11}$$

This simple equation can be used to predict the fraction of material that solidifies sufficiently fast to maintain the alignment of the LCP domains along the extrusion direction. Solidification requires that the temperature of the extruded material reaches the melting temperature ($T(x,t) = T_m$), whereas the preservation of flow-induced alignment is only possible if the cooling timescale t lies below the relaxation time of the LCP domains (τ). Using the skin depth λ to quantify the fraction of aligned material, we obtain the following relation:

$$\lambda = \frac{T_m - T_{RT}}{T_N - T_{RT}}\sqrt{\pi\alpha\tau} \tag{12}$$

The relaxation time τ provides a characteristic timescale for the exponential decay of the domain orientation parameter S, as described by the equation[36]:

$$S(t) = S_0 \exp\left(-\frac{t}{\tau}\right) \tag{13}$$

where S_0 is the initial flow-induced orientation. Because the orientation decays exponentially with time and minor misalignments are sufficient to strongly reduce the mechanical properties of the filament along the printing direction, we expect that cooling has to occur within timescales $t \ll \tau$ to obtain a filament shell with high mechanical properties.

Assuming that the core–shell structure is indeed controlled by thermal transport across the extrudate and that the mechanical properties of the filament are determined by the fraction of aligned LCP domains, we can use the above model to predict the dependence of the elastic modulus of the LCP filaments on the printing parameters. To this end, we estimate the elastic modulus E of the whole filament using a rule of mixture:

$$E = \phi_{skin}E_{skin} + \phi_{core}E_{core} \approx \phi_{skin}E_{skin} \tag{14}$$

Here E_{skin} is the modulus of the fully aligned polymer and ϕ_{skin} is the volume fraction of aligned material across the filament. ϕ_{skin} for horizontal and vertical filaments of diameter d_N can be geometrically described as:

$$\begin{gathered}\phi_{skin,V} = \frac{2\lambda}{d_N}\left(2 - \frac{2\lambda}{d_N}\right) \approx \frac{4\lambda}{d_N} \\ \phi_{skin,H} = 2\lambda\left(\frac{1}{d_N} + \frac{1}{h}\right)\end{gathered} \tag{15}$$

Introducing these relations into the rule-of-mixture equation yields the Young's modulus as a function of print parameters for the vertically and horizontally printed filaments:

$$\begin{gathered}E_V = \frac{4E_{skin}}{d_N}\sqrt{\pi\alpha\tau}\left(\frac{T_m - T_{RT}}{T_N - T_{RT}}\right) \\ E_H = \left(\frac{2}{d_N} + \frac{2}{h}\right)E_{skin}\sqrt{\pi\alpha\tau}\left(\frac{T_m - T_{RT}}{T_N - T_{RT}}\right)\end{gathered} \tag{16}$$

The experimentally measured Young's modulus of vertical filaments (Fig. 2e, g) was found to indeed scale with $1/d_N$ and $1/(T_N - T_{RT})$, as predicted by the equations proposed above. This supports the hypothesis that the mechanical properties of the filaments are controlled by the cooling kinetics of the extruded material.

This finding has direct implications for the 3D printing of tough LCP parts, because it implies that highly aligned filaments require not only the design of nozzles that promote shear and extensional flow, but also strategies to quickly extract heat from the extruded material. In our work, the shear and extensional flow was always above the critical value that is needed to orient the nematic domains of the LCP in the melt along the extrusion direction. The cooling was effectively achieved by depositing the material very close to the substrate or to the previously deposited (and cooled) polymer layers, such that thin print layers with low thermal diffusion length are obtained.

Classical laminate theory. The elastic modulus of printed LCP parts for distinct mechanical loading directions was theoretically estimated using classical laminate theory. To this end, we use a coordinate system in which direction 1 corresponds to the print direction and direction 2 lies perpendicular to it, both of which are orthogonal to direction 3. On the basis of this coordinate system, the following transformation matrix was applied:

$$\{\sigma\}_{12} = T\{\sigma\}_{xy}$$
$$T = \begin{bmatrix} \cos^2\theta & \sin^2\theta & 2\sin\theta\cos\theta \\ \sin^2\theta & \cos^2\theta & -2\sin\theta\cos\theta \\ -\sin\theta\cos\theta & \sin\theta\cos\theta & \cos^2\theta - \sin^2\theta \end{bmatrix} \qquad (17)$$

Using this matrix transformation, the Young's modulus in tensile direction x as a function of the print orientation angle θ can be calculated by the following equation[37]:

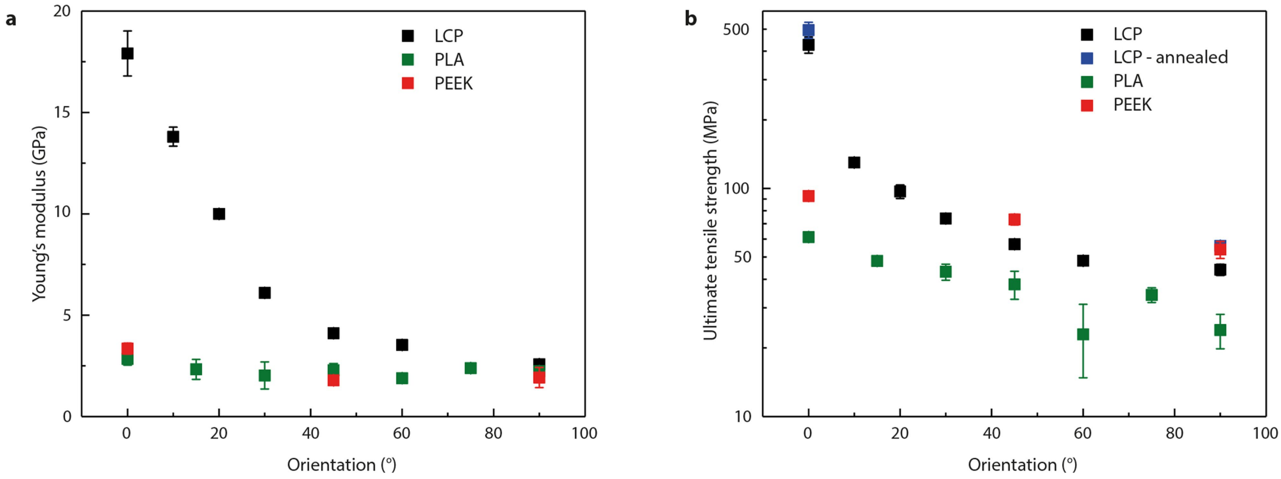
$$\frac{1}{E_x} = \frac{\cos^4\theta}{E_{11}} + \frac{\sin^4\theta}{E_{22}} + \left(\frac{1}{G_{12}} - \frac{2\nu_{12}}{E_{11}}\right)\cos^2\theta\sin^2\theta \qquad (18)$$

where E_{11} is the Young's modulus in the print direction, E_{22} is the modulus perpendicular to the print direction, G_{12} is the shear modulus and ν_{12} is Poisson's ratio. The material properties that were used for the calculations are listed in Extended Data Table 1 and were calculated from tensile tests.

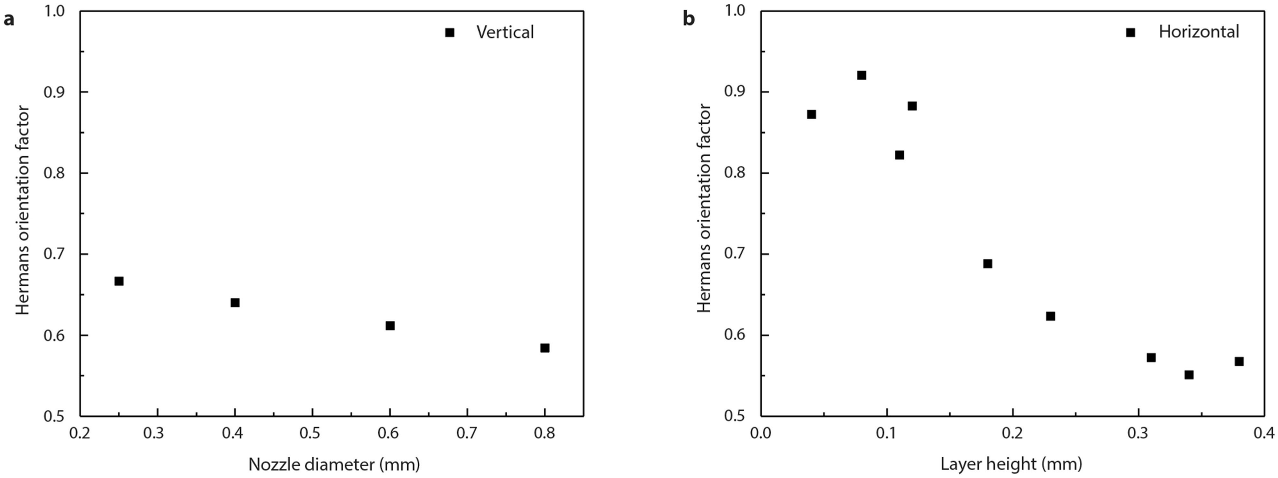
Code availability. The Grasshopper files for the generation of print paths are available from the corresponding authors on reasonable request.

Data availability. The datasets generated and analysed during this study are available from the corresponding authors on reasonable request. Source Data for Figs. 2–4 and Extended Data Figs. 1–6 and 8 are provided with the online version of the paper.

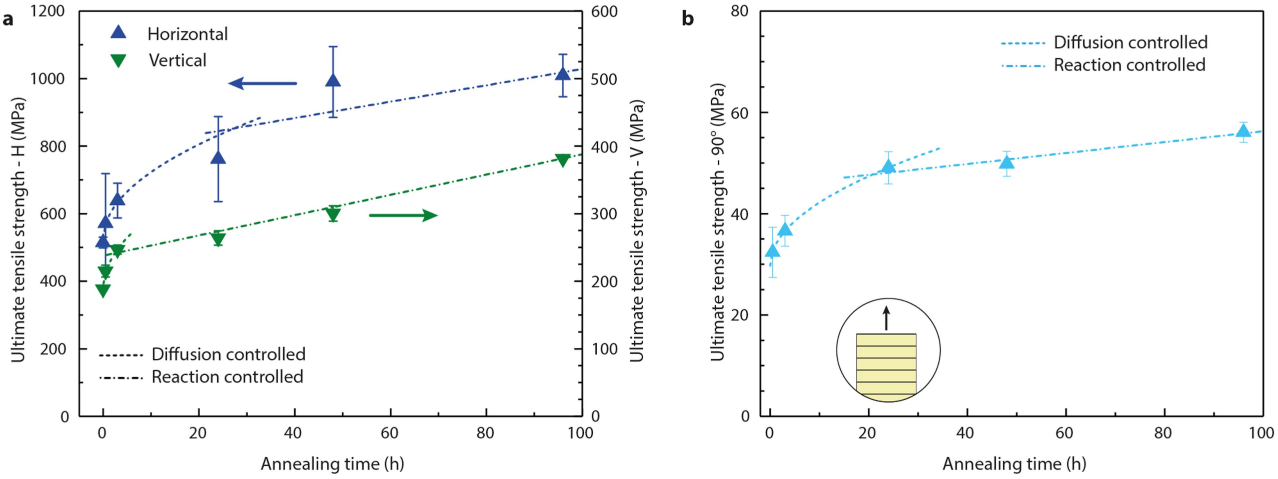
23. Sarlin, J. & Törmälä, P. Fiber formation and characterization of a thermotropic LCP. *J. Appl. Polym. Sci.* **40**, 453–469 (1990).
24. Chen, B.-K., Tsay, S.-Y. & Chen, J.-Y. Synthesis and properties of liquid crystalline polymers with low T_m and broad mesophase temperature ranges. *Polymer* **46**, 8624–8633 (2005).
25. Crevecoeur, G. & Groeninckx, G. Morphology and mechanical properties of thermoplastic composites containing a thermotropic liquid crystalline polymer. *Polym. Eng. Sci.* **30**, 532–542 (1990).
26. Motamedi, F., Jonas, U., Greiner, A. & Schmidt, H.-W. Preparation and characterization of fibres from a thermotropic liquid crystal polyester with non-coplanar biphenylene units. *Liq. Cryst.* **14**, 959–970 (1993).
27. Rendon, S., Burghardt, W. R. & Bubeck, R. A. Orientation dynamics in commercial thermotropic liquid crystalline polymers in transient shear flows. *Rheol. Acta* **46**, 945–956 (2007).
28. Warner, S. B. & Lee, J. Towards understanding the increase in strength of thermotropic polyesters with heat treatment. *J. Polym. Sci. B* **32**, 1759–1769 (1994).
29. Schindelin, J. et al. Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).
30. Dreher, S. et al. Chain orientation in melt-extruded samples of vectra A, vectra B, and blends in relation to the mechanical properties. *J. Appl. Polym. Sci.* **67**, 531–545 (1998).
31. Dickson, A. N., Barry, J. N., McDonnell, K. A. & Dowling, D. P. Fabrication of continuous carbon, glass and Kevlar fibre reinforced polymer composites using additive manufacturing. *Addit. Manuf.* **16**, 146–152 (2017).
32. Tymrak, B. M., Kreiger, M. & Pearce, J. M. Mechanical properties of components fabricated with open-source 3-D printers under realistic environmental conditions. *Mater. Des.* **58**, 242–246 (2014).
33. Kamthai, S. & Magaraphan, R. Thermal and mechanical properties of polylactic acid (PLA) and bagasse carboxymethyl cellulose (CMCB) composite by adding isosorbide diesters. *AIP Conf. Proc.* **1664**, 060006 (2015).
34. Arivazhagan, A. & Masood, S. Dynamic mechanical properties of ABS material processed by fused deposition modelling. *Int. J. Eng. Res. Appl.* **2**, 2013–2014 (2012).
35. Goyal, R. K., Tiwari, A. N., Mulik, U. P. & Negi, Y. S. Effect of aluminum nitride on thermomechanical properties of high performance PEEK. *Composites, Part A* **38**, 516–524 (2007).
36. Picken, S. J., Aerts, J., Visser, R. & Northolt, M. G. Structure and rheology of aramid solutions: X-ray scattering measurements. *Macromolecules* **23**, 3849–3854 (1990).
37. Gerdeen, J. C. & Rorrer, R. A. L. *Engineering Design with Polymers and Composites* 2nd edn, Ch. 8 (CRC Press, London, 2012).



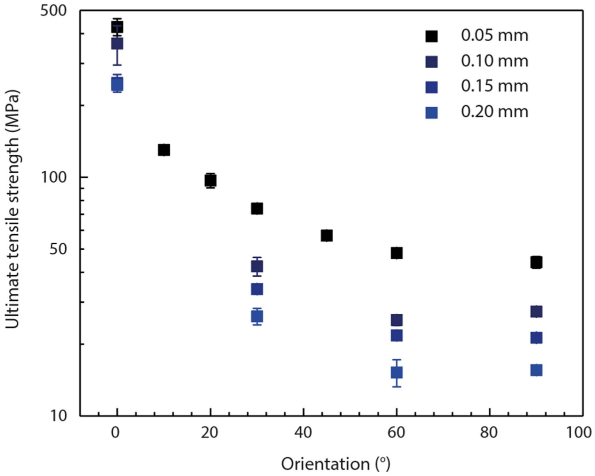
Extended Data Fig. 1 | Comparison between the mechanical properties of different FDM feedstock materials. a, Young's modulus as a function of print orientation for LCP, PLA and PEEK. **b**, Tensile strength of LCP, annealed LCP, PLA and PEEK. The strength of LCP samples is comparable to that of one of the strongest printable polymers (PEEK) at high orientation angles, and to our knowledge surpasses all other tested materials for print orientations below 20°. Error bars indicate the standard deviation for 5–10 measurements at each data point; see Source Data for details.



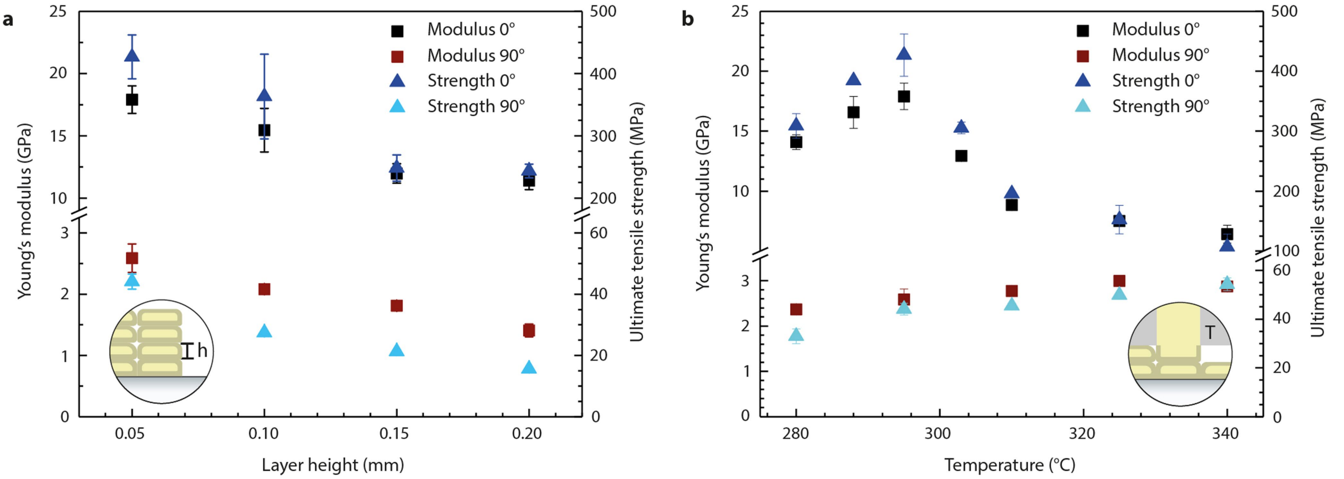
Extended Data Fig. 2 | Values of the Hermans' orientation factor for the LCPs. a, **b**, Hermans' orientation factors as a function of nozzle diameter (**a**) and layer height (**b**). Thinner samples have an increased skin-to-core material ratio and thus a higher orientation factor.



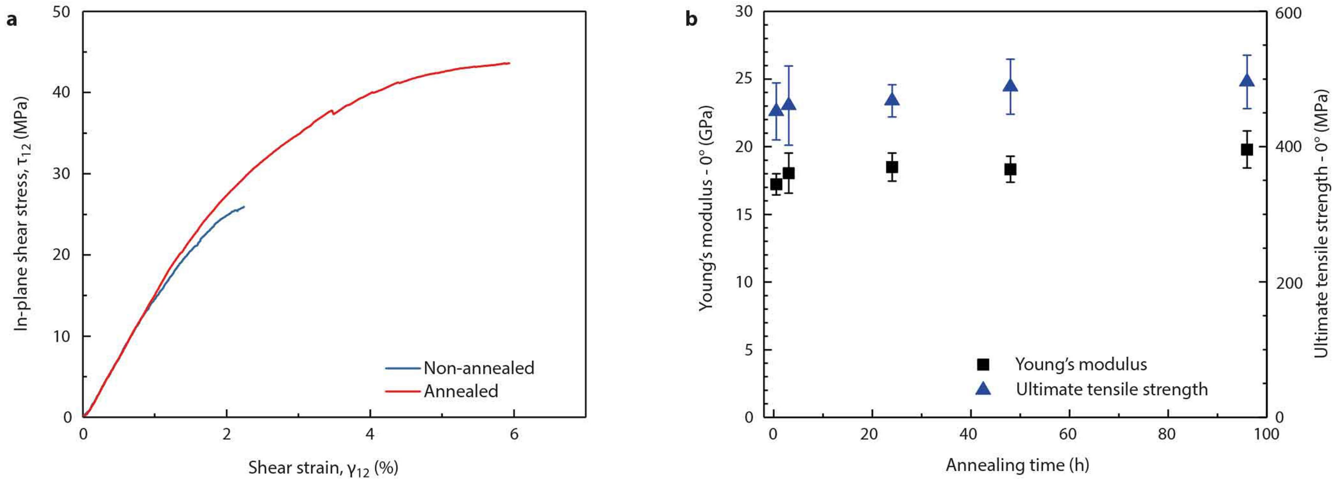
Extended Data Fig. 3 | Effect of annealing time on the ultimate strength of printed LCP filaments and parts. a, Strength (σ) as a function of annealing time (t) for single printed filaments. The experimental data suggest that the reaction is initially controlled by the diffusion of water out of the filament ($\sigma \propto t^{0.5}$) and later on by the intrinsic reaction kinetics ($\sigma \propto t$). **b**, Tensile strength as a function of annealing time for printed parts. The annealing kinetics follow the same trend as in the case of single filaments. Error bars indicate the standard deviation for 5–10 measurements at each data point; see Source Data for details.



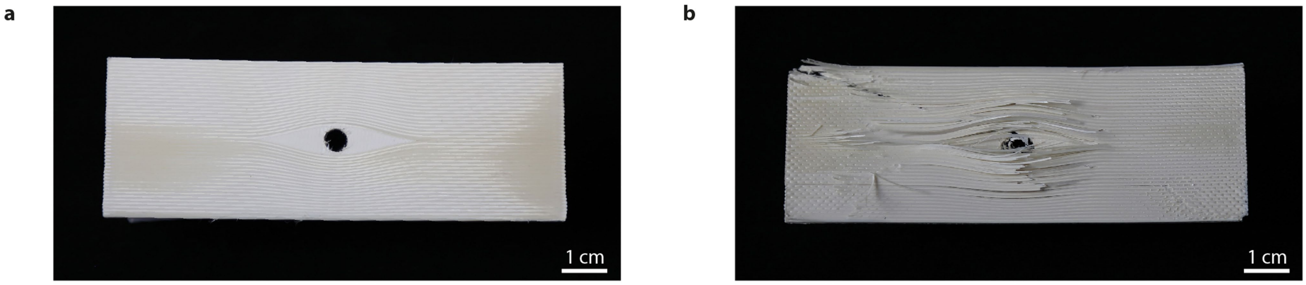
Extended Data Fig. 4 | Strength of LCP parts of different thicknesses as a function of print orientation. The observed decrease in tensile strength for increasing print orientations follows the decay in elastic modulus predicted from classical laminate theory (Fig. 3a). The lower strength at high orientation angles is also explained by a transition in failure mode from print filament fracture to interface fracture. Error bars indicate the standard deviation for 5–10 measurements at each data point; see Source Data for details.



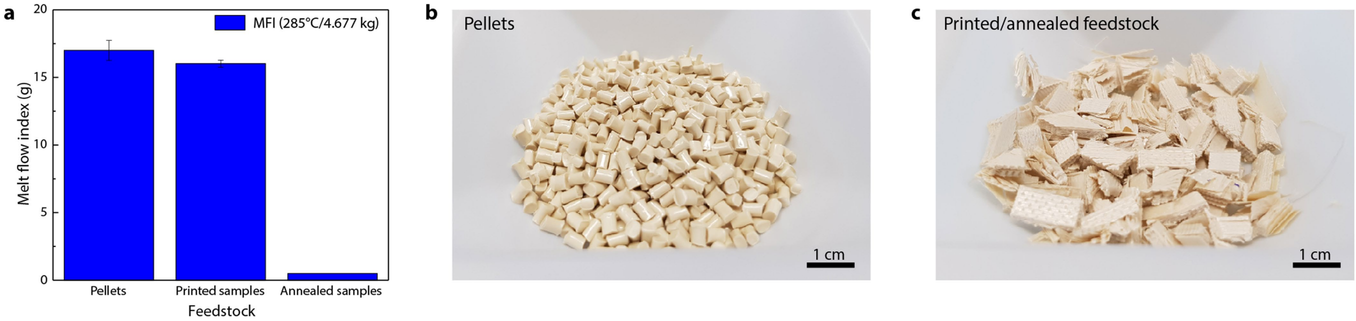
Extended Data Fig. 5 | Effect of print parameters on the Young's modulus of horizontally printed LCP parts. a, Tensile modulus and strength as a function of printed layer height. Lower layer heights lead to higher mechanical properties owing to the higher fraction of aligned skin relative to the material height. **b**, Tensile modulus and strength as a function of nozzle temperature (T_N). For temperatures higher than 300 °C, the tensile properties along the printing direction (0°) follows the behaviour observed for the Young's modulus of vertically printed filaments (Fig. 2g). The lower values obtained for the Young's modulus and strength at $T_N < 300$ °C probably result from poorer material flow and poorer print filament adhesion at lower nozzle temperatures. The tensile properties perpendicular to the print direction (90°) increase at higher temperatures, which inidicates an improved adhesion between hot printed filaments. Error bars indicate the standard deviation for 5–10 measurements at each data point; see Source Data for details.



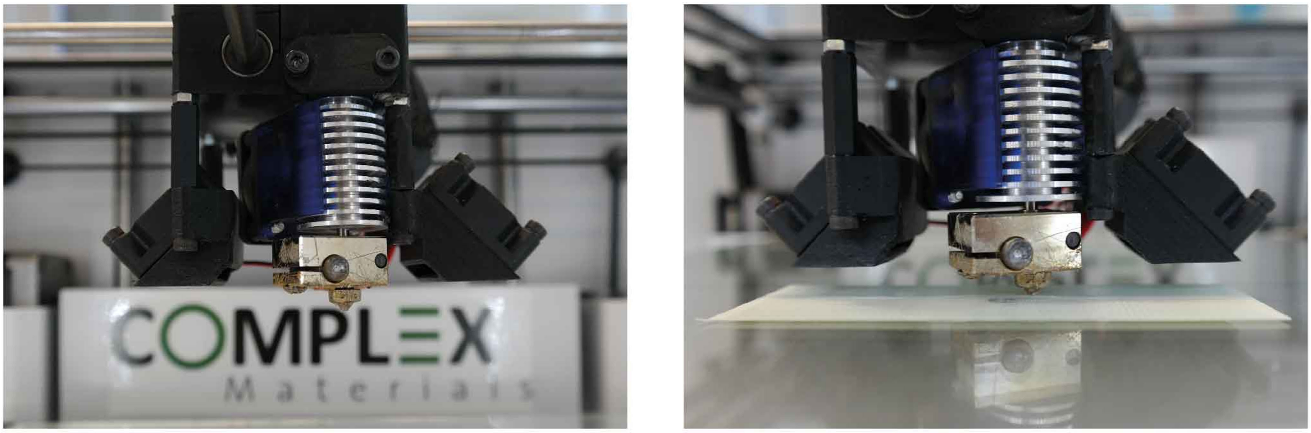
Extended Data Fig. 6 | Increased load transfer between filaments after annealing. a, Representative curve of shear stress versus strain shows that the shear strength of the LCP increases by 75% after thermal annealing for 96 h owing to the enhanced inter-filament adhesion. The shear strength data were obtained from ±45° tensile samples. **b**, The Young's modulus measured along the printing direction is found to slightly increase with longer annealing times. Error bars indicate the standard deviation for 5–10 measurements at each data point; see Source Data for details.



Extended Data Fig. 7 | Directional OHT sample. a, Print lines are guided around the hole, resembling a wood knot hole. The stress-shielded area next to the hole is reinforced with 90° filaments. **b**, The extensive damage to the same sample without catastrophic failure during tensile testing illustrates the high fracture toughness of the hierarchically structured architecture.



Extended Data Fig. 8 | Recyclability of printed LCP material. a, Melt flow index (MFI) of pristine pellets as well as recycled printed parts and recycled annealed printed parts (five measurements per material). The melt flow index quantifies the fluidity of the material at the indicated temperature and load. Higher indices indicate low viscosity. Non-annealed printed samples are readily recyclable owing to their high melt flow index. Although the higher molecular weight of annealed samples markedly reduces their fluidity at regular printing temperatures, this material is potentially recyclable if hydrolysis reactions are used to decrease its molecular weight and thus recover processability. Error bars indicate the standard deviation for 5 measurements at each data point; see Source Data for details. **b**, **c**, Examples of pellets (**b**) and recycled printed samples (**c**) as feedstock material.


COMPLEX
Materials

Extended Data Fig. 9 | Geared direct drive extruder with all-metal V6 hotend on an Ultimaker 2+.

Extended Data Table 1 | Classical laminate theory parameters used for print filament layer heights between 0.05 mm and 0.20 mm

Layer height (mm)	0.05	0.10	0.15	0.20
E_{11} (GPa)	17.90	15.40	11.97	11.39
E_{22} (GPa)	2.58	2.08	1.81	1.40
G_{12} (GPa)	1.70	1.30	1.10	1.00
ν_{12}	0.57	0.57	0.57	0.57

https://doi.org/10.1038/s41586-018-0476-5

Future response of global coastal wetlands to sea-level rise

Mark Schuerch[1,2]*, Tom Spencer[2], Stijn Temmerman[3], Matthew L. Kirwan[4], Claudia Wolff[5], Daniel Lincke[6], Chris J. McOwen[7], Mark D. Pickering[8], Ruth Reef[9], Athanasios T. Vafeidis[5], Jochen Hinkel[6,10], Robert J. Nicholls[11] & Sally Brown[11]

The response of coastal wetlands to sea-level rise during the twenty-first century remains uncertain. Global-scale projections suggest that between 20 and 90 per cent (for low and high sea-level rise scenarios, respectively) of the present-day coastal wetland area will be lost, which will in turn result in the loss of biodiversity and highly valued ecosystem services[1–3]. These projections do not necessarily take into account all essential geomorphological[4–7] and socio-economic system feedbacks[8]. Here we present an integrated global modelling approach that considers both the ability of coastal wetlands to build up vertically by sediment accretion, and the accommodation space, namely, the vertical and lateral space available for fine sediments to accumulate and be colonized by wetland vegetation. We use this approach to assess global-scale changes in coastal wetland area in response to global sea-level rise and anthropogenic coastal occupation during the twenty-first century. On the basis of our simulations, we find that, globally, rather than losses, wetland gains of up to 60 per cent of the current area are possible, if more than 37 per cent (our upper estimate for current accommodation space) of coastal wetlands have sufficient accommodation space, and sediment supply remains at present levels. In contrast to previous studies[1–3], we project that until 2100, the loss of global coastal wetland area will range between 0 and 30 per cent, assuming no further accommodation space in addition to current levels. Our simulations suggest that the resilience of global wetlands is primarily driven by the availability of accommodation space, which is strongly influenced by the building of anthropogenic infrastructure in the coastal zone and such infrastructure is expected to change over the twenty-first century. Rather than being an inevitable consequence of global sea-level rise, our findings indicate that large-scale loss of coastal wetlands might be avoidable, if sufficient additional accommodation space can be created through careful nature-based adaptation solutions to coastal management.

Coastal wetlands provide many important ecosystem services (valued up to US\$194,000 $\mathrm{ha}^{-1}\,\mathrm{yr}^{-1}$)[9], including carbon sequestration[10,11], natural coastal protection[12–15], support of fisheries[16] and water quality improvement[17]. Recent global-scale assessments of coastal wetland dynamics have suggested that the ability of many marshes and mangroves to build up vertically has already been overwhelmed by present-day sea-level rise (SLR), leading to widespread wetland loss[1–3]. At the same time, more regional to local-scale field measurements and models of salt marsh accretion have concluded that most large-scale assessments have overestimated the vulnerability of coastal wetlands to SLR[4]. These differences highlight a major knowledge gap in our understanding of the responses of coastal wetland areas to global environmental change. It has been argued that the reason for the observed discrepancy is that large-scale assessments have so far failed to consider the well understood biophysical feedback mechanisms that are typically included in local-scale models[4]. These mechanisms include the ability of coastal wetlands to build up vertically by sediment accretion, which is enhanced with increasing inundation heights and frequencies, triggered, for example, by accelerating SLR, and which enables coastal wetlands to persist or even prosper with SLR[5–7].

A second limitation of previous global-scale assessments is that they have not yet represented accommodation space (the vertical and lateral space available for fine sediments to accumulate and be colonized by wetland vegetation) in a spatially explicit manner[2,4]. This constitutes an important gap, as recent papers have suggested that anthropogenic barriers to inland wetland migration (such as coastal flood protection structures, coastal roads and railway lines, settlements, and impervious land surfaces) may be a more important threat to coastal wetlands than drowning by SLR alone[2,4,18].

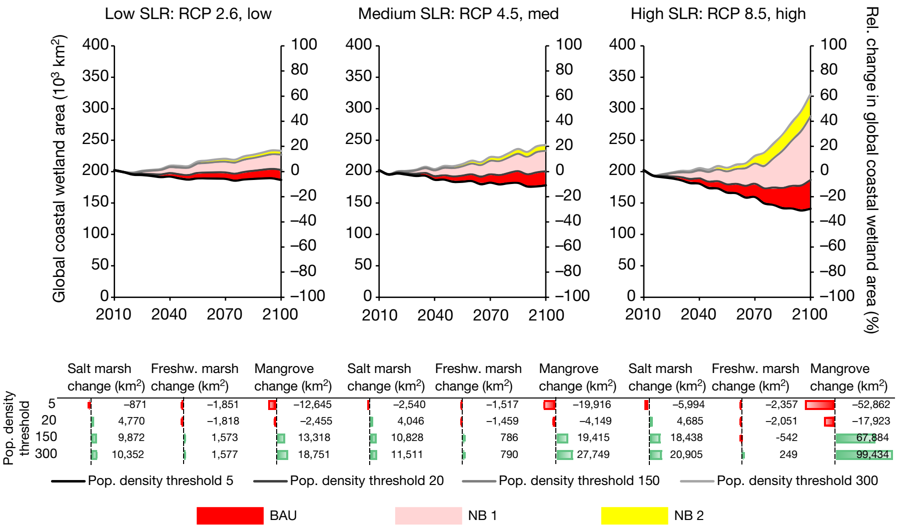
We address both of these limitations, and assess global-scale changes in coastal wetland areas in response to global SLR and anthropogenic coastal occupation, using a new integrated modelling approach. We consider (1) the vertical adaptability of coastal wetlands by biophysical feedbacks between wetland accretion and SLR, assuming current-day levels of sediment availability, and (2) their horizontal adaptability, as determined by the interactions between inland wetland migration and anthropogenic barriers, assuming wetland inland migration to be a function of accommodation space[8]. We present a model to make projections of the global resilience of coastal wetlands to twenty-first century SLR scenarios under existing and increased accommodation space, representing present conditions and two additional coastal management scenarios following the wider implementation of nature-based adaptation strategies[12]. By means of a comprehensive sensitivity analysis, we finally assess the extent to which this resilience is controlled by vertical and horizontal adaptation mechanisms.

On the basis of the simulation runs during model calibration, our calibrated model, which includes mangroves as well as tidal salt and freshwater marshes, correctly predicts observations of present-day vertical wetland change, obtained from large meta-datasets from all over the world[3,4,19], for 78% of all coastal areas where data are currently available ($N = 46$) (Extended Data Table 1, Extended Data Fig. 1). Although our model performs very well in regions where coastal wetlands were reported to be stable (that is, with vertical wetland growth in balance with local SLR) or drowning (that is, slower vertical wetland growth than local SLR), it tends to underestimate the number of locations with an elevation surplus (that is, faster vertical wetland growth than local SLR). Hence, our predictions of the ability of wetlands to vertically grow in pace with twenty-first century SLR rates may be considered conservative.

Projections of the extent of coastal wetlands by 2100 are based on simulations using three different regionalized relative SLR scenarios (Representative Concentration Pathways (RCPs) 2.6, 4.5 and 8.5,

---

[1]Lincoln Centre for Water and Planetary Health, School of Geography, University of Lincoln, Lincoln, UK. [2]Cambridge Coastal Research Unit, Department of Geography, University of Cambridge, Cambridge, UK. [3]Ecosystem Management Research Group, University of Antwerp, Antwerp, Belgium. [4]Virginia Institute of Marine Science, College of William and Mary, Gloucester Point, VA, USA. [5]Institute of Geography, Christian-Albrechts University of Kiel, Kiel, Germany. [6]Global Climate Forum, Berlin, Germany. [7]UN Environment World Conservation Monitoring Centre, Cambridge, UK. [8]Ocean and Earth Science, National Oceanography Centre, University of Southampton, Southampton, UK. [9]School of Earth, Atmosphere and Environment, Monash University, Clayton, Victoria, Australia. [10]Division of Resource Economics, Thaer-Institute and Berlin Workshop in Institutional Analysis of Social-Ecological Systems (WINS), Humboldt-University, Berlin, Germany. [11]Faculty of Engineering and the Environment, University of Southampton, Southampton, UK. *e-mail: mschuerch@lincoln.ac.uk



Pop. density threshold	Salt marsh change (km^2)	Freshw. marsh change (km^2)	Mangrove change (km^2)	Salt marsh change (km^2)	Freshw. marsh change (km^2)	Mangrove change (km^2)	Salt marsh change (km^2)	Freshw. marsh change (km^2)	Mangrove change (km^2)
5	–871	–1,851	–12,645	–2,540	–1,517	–19,916	–5,994	–2,357	–52,862
20	4,770	–1,818	–2,455	4,046	–1,459	–4,149	4,685	–2,051	–17,923
150	9,872	1,573	13,318	10,828	786	19,415	18,438	-542	67,884
300	10,352	1,577	18,751	11,511	790	27,749	20,905	249	99,434

Fig. 1 | Global change in coastal wetland areas. Results are displayed for all three SLR scenarios (RCP 2.6, low; RCP 4.5, medium; RCP 8.5, high) and three human adaptation scenarios, defined by different population density thresholds (BAU 1: 5–20 people km^{-2}, NB 1: 20–150 people km^{-2}, NB 2: 150–300 people km^{-2}). Sediment accretion is considered, and wetland inland migration is limited to where the population density in the 1-in-100-year floodplain falls below the respective threshold. Areal changes of all three wetland types (salt marsh, freshwater marsh and mangrove) are indicated in the tables below the graphs.

corresponding to an SLR of 29, 50 and 110 cm, respectively, by 2100) and three human adaptation scenarios with varying degrees of available accommodation space (Extended Data Table 2): (i) a business-as-usual (BAU) scenario, in which we assume that no accommodation space is available where local population densities in the 1-in-100-year coastal floodplain exceed thresholds between 5 and 20 people km^{-2}; (ii) a moderate level of nature-based adaptation (NB 1), in which the population density threshold ranges between 20 and 150 people km^{-2}; and (iii) a high level of nature-based adaptation (NB 2), with population density thresholds between 150 and 300 people km^{-2}. Changes in population growth during the simulation period are considered by applying a scenario of national population growth rates based on the Shared Socio-economic Pathway SSP2 (IIASA)[20], which is characterized by a moderate, and after 2070 slowing, global population growth, leading to 9 billion people by 2100[21].

Under all SLR scenarios, 20 people km^{-2} constitutes a crucial population density threshold. If a higher population density threshold is applied, more coastal wetlands have sufficient accommodation space to migrate inland resulting in an overall gain in global coastal wetland area (Fig. 1). If lower thresholds are considered, less coastal wetlands have sufficient accommodation space resulting in an overall global loss. The population density threshold of 20 people km^{-2} corresponds to what we estimate as the current global average above which coastal communities are protected by some type of coastal protection infrastructure (Supplementary Information), hence allowing inland migration for only 37% of all global coastal wetlands. A population density threshold of 300 people km^{-2} is the lower threshold for urban developments, as defined by the European Commission[22], and sets the upper limit for potential wetland inland migration (NB 2 scenario). The highest SLR scenario at this threshold results in a substantial increase in global coastal wetland area (+60%). The same SLR scenario with a threshold population density of 5 people km^{-2} results in a net global loss of 30% (Fig. 1). When applying the lowest SLR scenario, areal coastal wetland changes for population density thresholds between 5 and 300 people km^{-2} only range between −8% (loss) and +15% (gain) (Fig. 1). The largest changes are observed for mangroves, which make the largest contribution to the global wetland area from the beginning (69%). Notably, much smaller losses are observed for salt marshes, even under the human adaptation scenarios with the least accommodation space (Fig. 1).

Under the BAU scenario for accommodation space (5–20 people km^{-2}), changes in the extent of global coastal wetlands range between −8% (loss) and 0% (no change) for the lowest SLR scenario and between −30% (loss) and −8% (loss) for the highest SLR scenario. These losses can primarily be attributed to an increasing sediment deficiency, impeding the ability of the wetland to keep pace vertically with SLR. If, coastal wetlands are given more accommodation space in the future (for example, in the context of the implementation of nature-based adaptation solutions), global coastal wetlands could increase in areal extent (Fig. 1). Our moderate nature-based adaptation scenario (NB 1: 20–150 people km^{-2}) results in an increase between 0% and 12% for the low, and between −8% (loss) and 42% for the high, SLR scenario. Under the more extreme adaptation scenario (NB 2: 150–300 people km^{-2}), we anticipate even higher increases, between 12% and 15% for the low, and between 42% and 60% for the high, SLR scenario (Fig. 1). In contrast to the BAU scenario, these gains for the moderate and extreme nature-based adaptation scenarios (NB 1 and NB 2) are driven by inland wetland migration rather than vertical sediment accretion, and therefore independent of sediment availability.

Under the BAU scenario (lower boundary: 5 people km^{-2}), most of the absolute loss in coastal wetland areas (approximately 66%) is projected to occur in the Caribbean Sea, the southern US east coast and parts of southeast Asia (Fig. 2a). Similarly, southeast Asia was previously identified[19] as a highly critical region for mangrove resilience to SLR. The patterns of expected relative changes in wetland areas (that is, the percentage gain or loss) are different, but essentially confirm the Dynamic Interactive Vulnerability Assessment (DIVA) modelling results[2]; the largest relative area losses (again, under a scenario of highly constrained accommodation space) are found in the Caribbean Sea, along the eastern US coast and in the western Baltic Sea, the Mediterranean Sea, the Red Sea and parts of southeast Asia (Fig. 2b).

The spatial patterns of coastal wetland loss strongly resemble those of the modelled present-day sediment balance, namely the difference between the sediment required for a coastal wetland surface to keep pace vertically with current local relative SLR and the current-day sediment availability (Fig. 3). For example, large regions of sediment deficit are identified in the Caribbean Sea, western Baltic Sea, Mediterranean Sea, and along the eastern and western US coasts (Fig. 3). These areas largely coincide with the hotspot regions for relative wetland area losses under a scenario of highly constrained accommodation space (Fig. 2b). Meanwhile, most parts of Asia, South America and northwest Europe show sufficient or excess sediment availability (Fig. 3), which correspond to areas with small relative wetland loss, even where accommodation space is limited, as vertical sediment accretion counteracts relative SLR (Fig. 2b).

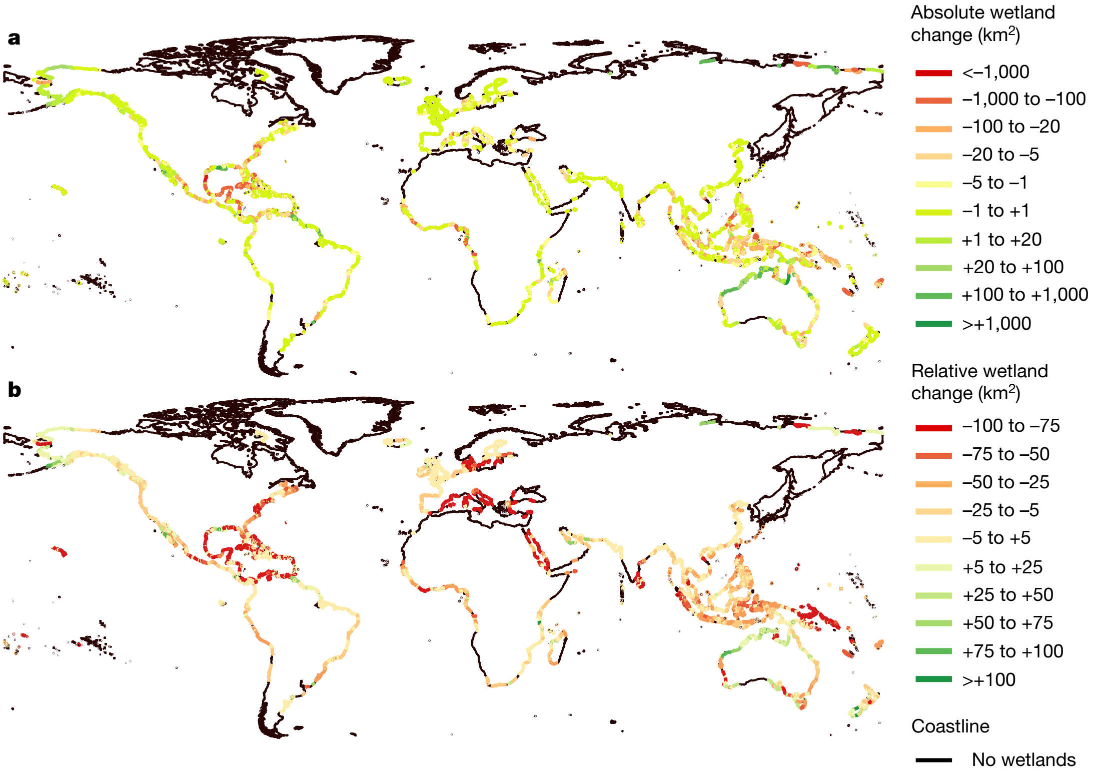

a
b
Absolute wetland change (km^2)
<–1,000
–1,000 to –100
–100 to –20
–20 to –5
–5 to –1
–1 to +1
+1 to +20
+20 to +100
+100 to +1,000
>+1,000
Relative wetland change (km^2)
–100 to –75
–75 to –50
–50 to –25
–25 to –5
–5 to +5
+5 to +25
+25 to +50
+50 to +75
+75 to +100
>+100
Coastline
No wetlands

Fig. 2 | Spatial distribution of coastal wetland change. **a**, **b**, Absolute (**a**) and relative (**b**) changes in coastal wetland areas are displayed for the medium SLR scenario (RCP 4.5), assuming inhibition of wetland inland migration everywhere, but in (nearly) uninhabited regions with a population density of less than 5 people km^{-2}. Population density is subject the population growth throughout the simulation period, following SSP2[20,21]. The displayed coastline was generated during the DINAS-COAST FP5-EESD EU project (EVK2-CT-2000-00084).

Our sensitivity analysis confirms the importance of accounting for vertical sediment accretion with our 'sediment accretion only' scenario (hypothetical scenario (HYS) 2, Extended Data Table 2). This scenario reduces the global loss of coastal wetlands from 38% to 20%, 50% to 26% and 77% to 54% for the low, medium and high SLR scenarios, respectively, as compared to our 'no resilience' scenario, in which no accommodation space and no vertical sediment accretion are assumed (HYS 4; Extended Data Table 2, Extended Data Fig. 2).

Previous studies have highlighted the dangers of low sediment availability and reduced sediment supply, which may be exacerbated regionally by increasing the numbers of dams being built within river catchments, causing increased risk for coastal wetland loss with SLR[23–25]. However, our model sensitivity analysis under the high SLR scenario (RCP 8.5), and accounting for vertical sediment accretion, demonstrates that if present-day values of sediment supply were to change by ±50%, only a ±6% change in global wetland area would result (Extended Data Table 3). By contrast, accommodation space for inland wetland migration has a much stronger control on wetland persistence with SLR, yet much less is known about the actual process and further research is urgently needed. Our sensitivity analysis shows that even in heavily sediment-starved regions, an increase in accommodation space could result in a net wetland gain (Extended Data Fig. 3), particularly under high rates of SLR, even though the seaward side of the wetland could regularly be lost owing to the lack of sediment. Under

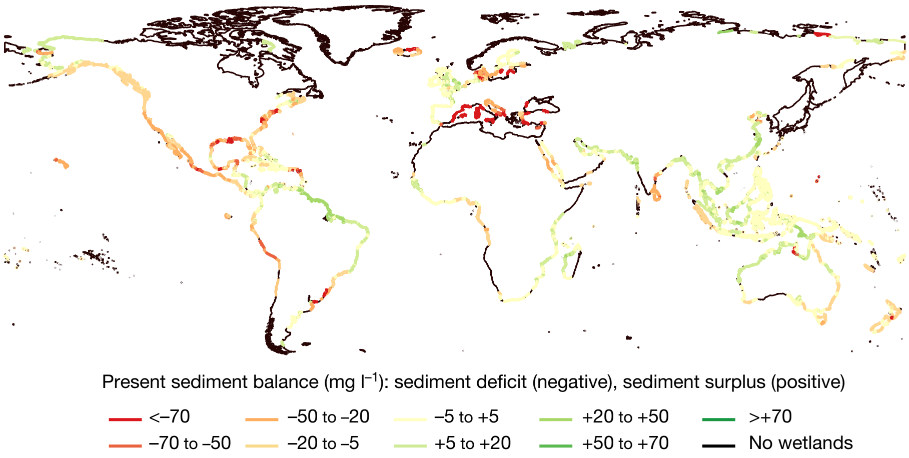

Present sediment balance (mg l^{-1}): sediment deficit (negative), sediment surplus (positive)
<–70
–70 to –50
–50 to –20
–20 to –5
–5 to +5
+5 to +20
+20 to +50
+50 to +70
>+70
No wetlands

Fig. 3 | Present-day global sediment balance. Sediment surplus (positive values) or sediment deficit (negative values) (in mg l^{-1}) represent the difference between the suspended sediment concentration needed for coastal wetlands to build up vertically with current SLR rates and the actual total suspended matter concentration derived from the satellite-borne GlobColour data (http://globcolour.info). The displayed coastline was generated during the DINAS-COAST FP5-EESD EU project (EVK2-CT-2000-00084).

extreme rates of SLR, and where sediment availability is insufficient, future coastal wetlands may therefore have a shorter lifetime and a lower degree of geomorphological, hydrological and biogeochemical complexity[26].

It should be noted that locally and especially in delta regions, these global mechanisms may not be as straightforward because historical and contemporary catchment and delta practices (for example, river damming and dredging) are responsible for much of the observed coastal wetland trends in many 'loss hotspots' rather than global SLR[27]. Also, constraints on the inland migration of coastal wetlands may arise from adverse soil conditions, particularly where the inundated land has been intensively modified by humans, unsuitable geomorphological characteristics or elevation constraints (if located too low in the tidal frame)[26,28]. To alleviate these constraints, coastal management strategies and engineering may locally be required to facilitate coastal wetlands to migrate inland[26]. As a consequence, local patterns of wetland resilience may be at considerable variance with global estimates of change.

Our model projections suggest that nature-based adaptation solutions that maximize the inland migration of tidal wetlands in response to SRL, wherever possible, may help safeguard wetland persistence with SLR and protect associated ecosystem services. Existing nature-based adaptation solutions that allow coastal wetlands to migrate inland include the inland displacement of coastal flood defences (typically along highly engineered coastlines)[12] or the designation of nature reserve buffers in upland areas surrounding coastal wetlands[18]. These schemes, however, are currently implemented as local-scale projects only; strategically upscaling such projects, for example, as suggested by the shoreline management plans in England and Wales[29] or the coastal master plan in Lousiana[30], may help coastal wetlands adapt to SLR at the landscape scale and protect rapidly increasing global coastal populations.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0476-5

Received: 2 September 2017; Accepted: 25 July 2018;
Published online 12 September 2018.

1. Blankespoor, B., Dasgupta, S. & Laplante, B. Sea-level rise and coastal wetlands. *Ambio* **43**, 996–1005 (2014).
2. Spencer, T. et al. Global coastal wetland change under sea-level rise and related stresses: the DIVA Wetland Change Model. *Global Planet. Change* **139**, 15–30 (2016).
3. Crosby, S. C. et al. Salt marsh persistence is threatened by predicted sea-level rise. *Estuar. Coast. Shelf Sci.* **181**, 93–99 (2016).
4. Kirwan, M. L., Temmerman, S., Skeehan, E. E., Guntenspergen, G. R. & Fagherazzi, S. Overestimation of marsh vulnerability to sea level rise. *Nat. Clim. Chang.* **6**, 253–260 (2016).
5. Schuerch, M., Vafeidis, A., Slawig, T. & Temmerman, S. Modeling the influence of changing storm patterns on the ability of a salt marsh to keep pace with sea level rise. *J. Geophys. Res. Earth Surf.* **118**, 84–96 (2013).
6. French, J. R. Numerical simulation of vertical marsh growth and adjustment to accelerated sea-level rise, North Norfolk, U.K. *Earth Surf. Process. Landf.* **18**, 63–81 (1993).
7. Morris, J. T., Sundareshwar, P. V., Nietch, C. T., Kjerfve, B. & Cahoon, D. R. Responses of coastal wetlands to rising sea level. *Ecology* **83**, 2869–2877 (2002).
8. Enwright, N. M., Griffith, K. T. & Osland, M. J. Barriers to and opportunities for landward migration of coastal wetlands with sea-level rise. *Front. Ecol. Environ.* **14**, 307–316 (2016).
9. Costanza, R. et al. Changes in the global value of ecosystem services. *Glob. Environ. Change* **26**, 152–158 (2014).
10. Duarte, C. M., Losada, I. J., Hendriks, I. E., Mazarrasa, I. & Marba, N. The role of coastal plant communities for climate change mitigation and adaptation. *Nat. Clim. Chang.* **3**, 961–968 (2013).
11. McLeod, E. et al. A blueprint for blue carbon: toward an improved understanding of the role of vegetated coastal habitats in sequestering CO_2. *Front. Ecol. Environ.* **9**, 552–560 (2011).
12. Temmerman, S. et al. Ecosystem-based coastal defence in the face of global change. *Nature* **504**, 79–83 (2013).
13. Möller, I. et al. Wave attenuation over coastal salt marshes under storm surge conditions. *Nat. Geosci.* **7**, 727–731 (2014).
14. Shepard, C. C., Crain, C. M. & Beck, M. W. The protective role of coastal marshes: a systematic review and meta-analysis. *PLoS One* **6**, e27374 (2011).
15. Stark, J., Van Oyen, T., Meire, P. & Temmerman, S. Observations of tidal and storm surge attenuation in a large tidal marsh. *Limnol. Oceanogr.* **60**, 1371–1381 (2015).
16. Aburto-Oropeza, O. et al. Mangroves in the Gulf of California increase fishery yields. *Proc. Natl Acad. Sci. USA* **105**, 10456–10459 (2008).
17. Teuchies, J. et al. Estuaries as filters: the role of tidal marshes in trace metal removal. *PLoS One* **8**, e70381 (2013).
18. Kirwan, M. L. & Megonigal, J. P. Tidal wetland stability in the face of human impacts and sea-level rise. *Nature* **504**, 53–60 (2013).
19. Lovelock, C. E. et al. The vulnerability of Indo-Pacific mangrove forests to sea level rise. *Nature* **526**, 559–563 (2015).
20. van Vuuren, D. P. et al. A new scenario framework for climate change research: scenario matrix architecture. *Clim. Change* **122**, 373–386 (2014).
21. Kc, S. & Lutz, W. The human core of the shared socioeconomic pathways: population scenarios by age, sex and level of education for all countries to 2100. *Glob. Environ. Change* **42**, 181–192 (2017).
22. Dijkstra, L. & Poelman, H. *A Harmonised Definition of Cities and Rural Areas: the New Degree of Urbanisation*; European Commission Working Paper WP 01/2014 (2014).
23. Day, J. W., Pont, D., Hensel, P. F. & Ibàñez, C. Impacts of sea-level rise on deltas in the Gulf of Mexico and the Mediterranean: the importance of pulsing events to sustainability. *Estuaries* **18**, 636–647 (1995).
24. Yang, S. L. et al. Impact of dams on Yangtze River sediment supply to the sea and delta intertidal wetland response. *J. Geophys. Res. Earth Surf.* **110**, F03006 (2005).
25. Ganju, N. K. et al. Spatially integrative metrics reveal hidden vulnerability of microtidal salt marshes. *Nat. Commun.* **8**, 14156 (2017).
26. Spencer, K. L. et al. Physicochemical changes in sediments at Orplands Farm, Essex, UK following 8 years of managed realignment. *Estuar. Coast. Shelf Sci.* **76**, 608–619 (2008).
27. Jankowski, K., Törnqvist, T. E. & Fernandes, A. M. Vulnerability of Louisiana's coastal wetlands to present-day rates of relative sea-level rise. *Nat. Commun.* **8**, 14792 (2017).
28. French, P. W. Managed realignment — the developing story of a comparatively new approach to soft engineering. *Estuar. Coast. Shelf Sci.* **67**, 409–423 (2006).
29. Nicholls, R. J., Townend, I. H., Bradbury, A. P., Ramsbottom, D. & Day, S. A. Planning for long-term coastal change: experiences from England and Wales. *Ocean Eng.* **71**, 3–16 (2013).
30. Peyronnin, N. et al. Louisiana's 2012 coastal master plan: overview of a science-based and publicly informed decision-making process. *J. Coast. Res.* **67**, 1–15 (2013).

Acknowledgements This research was financially supported by the Deutsche Forschungsgemeinschaft (DFG) through the Cluster of Excellence 80 'The Future Ocean', funded within the framework of the Excellence Initiative on behalf of the German federal and state governments, the personal research fellowship of M.S. (project number 272052902) and by the Cambridge Coastal Research Unit (Visiting Scholar Programme). Furthermore, this work has partly been supported by the European Union's Seventh Programme for Research, Technological Development and Demonstration (grant no. 603396, RISES-AM project), the European Union's Horizon 2020 Research and Innovation Programme (grant no. 642018, GREEN-WIN project), the US National Science Foundation (Coastal SEES 1426981 and NSF CAREER 1654374), Deltares and the UK Natural Environment Research Council. We thank M. Martin for support in editing the calibration data and G. Amable for statistical advice.

Reviewer information *Nature* thanks B. Glavovic, J. Woodruff and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions M.S. and T.S. developed the model algorithm. M.S. and D.L. developed the model code. M.S., C.W., C.McO., M.D.P., M.L.K., A.T.V., R.R. and S.B. gathered and produced input data. M.S., S.T. and R.J.N. analysed and interpreted the model simulations. M.S., T.S., S.T. M.L.K. and J.H. wrote the paper.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0476-5.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0476-5.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to M.S.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

General description of model approach. Our model is based on the construction of coastal profiles for 12,148 coastline segments. These segments constitute the spatial units of the Dynamic Interactive Vulnerability Assessment (DIVA) modelling framework[31,32]. The coastal profiles are derived from the Shuttle Radar Topography Mission (SRTM) floodplain data, available from the global DIVA database[33]. Within each coastline segment, the existing coastal wetlands, as reported by the United Nations Environment Programme World Conservation Monitoring Centre (UNEP WCMC)[34,35], are assumed to be located between mean sea level (MSL) and mean high water spring (MHWS) level. With SLR, the seaward side of the wetlands are increasingly inundated ('unconstrained wetland loss'), while the landward side migrates inland by converting terrestrial uplands to coastal wetlands[36] (Extended Data Fig. 5). However, inland wetland migration may be inhibited by anthropogenic coastal infrastructure that reduces the available accommodation space[36–39], a variable that we approximate with the population density in the floodplain of the 1-in-100-year extreme water level (Extended Data Fig. 4).

Seaward wetland loss through inundation is counteracted by a large tidal range and a high sediment availability, as both these variables increase the resilience of coastal wetlands towards drowning through vertical sediment accretion processes[19,40–44]. This is represented by the wetland adaptability score (WAS) reducing the loss of wetlands where tidal range and sediment availability are high[40] (Extended Data Fig. 4). The calculation of the WAS is based on a linear relationship between sediment availability and wetland drowning, whereas the slope of the linear relationship depends on tidal range. This relationship was suggested previously[40], when an ensemble of five different tidal marsh accretion models was used to identify the critical rates of relative SLR as a function of tidal range and sediment availability.

Following the calculation of the seaward wetland loss and inland wetland gain, the resulting global coastal wetland areas are calculated for every model time step (5 years) between 2010 and 2100. The model is driven by temporal changes in the model variables 'regional relative sea level rise' and 'population density' according to a range of regionalized scenarios for global SLR (RCPs)[45] and SSP2[20] for national population growth, respectively (Extended Data Table 2, Extended Data Fig. 4).

Database and data model. The input variables are derived from spatially explicit global data sets. They are attributed to the 12,148 coastline segments[31], which have an average length of 57 km. Coastline segmentation is a product of the DIVA modelling framework; the related database includes more than 100 bio-physical and socio-economic parameters[31]. The dissection of the global coastline into segments is based on the concept described previously[46], where coastal units have been created such that bio-physical and socio-economic impacts of global SLR are expected to be comparable within each coastline segment.

Construction of the coastal topographic profile. For each of the DIVA coastline segments, the coastal topographical profile is approximated using the areal information on coastal floodplains as described previously[32]. Floodplain areas (km^2) are provided for the elevation increments <1.5 m, 1.5–2.5 m, 2.5–3.5 m, 3.5–4.5 m, 4.5–5.5 m, 5.5–8.5 m, 8.5–12.5 m and 12.5–16.5 m, based on freely available SRTM data[47]. The SRTM data have a 90 m horizontal and a 1 m vertical resolution. The coastal profiles are constructed by dividing the floodplain areas per elevation increment by the length of the corresponding coastline segment to calculate the inundation lengths, which are then plotted against the upper boundaries of the elevation increments (that is, 1.5 m, 2.5 m, 3.5 m and so on) (Extended Data Fig. 5). It is thereby assumed that elevations continuously increase with distance from the coast, which has been shown to be a reasonable assumption[33].

Elevations between the upper boundaries of the elevation increments are linearly interpolated following earlier global assessments[32,48–50]. Linear interpolation between the MHWS level and an elevation of 1.5 m (or higher) was previously shown[51,52] to approximate high resolution LIDAR-derived elevations with a mean error of less than 30 cm and to produce no systematic bias with respect to the area of inundated land, even for the lowest 50 cm of the profile[52].

Wetland data. The areal wetland extents used in the context of this study include current wetland areas (1973–2015) for mangrove forests[34], salt marshes[35] and tidal freshwater marshes[53]. On the basis of a literature search for the lower and upper elevation limits of mangroves, salt marshes and tidal freshwater marshes[54–58], we assume that all coastal wetland types are located at elevations between MSL and MHWS and can occur over the entire elevation range. The reported wetland areas for each coastline segment are distributed alongside the non-wetland floodplain on the previously constructed coastal profile (Extended Data Fig. 5). We appreciate that in nature, the upper and lower boundaries of coastal wetlands will vary as a result of different vegetation species, tidal currents and waves[59], but for our global application MSL as the lower, and MHWS as the upper, limit constitute solid boundaries.

Regional relative sea-level rise data and scenarios. We use three SLR scenarios, covering the range of global SLR as projected by the IPCC AR5[45] plus a possible greater contribution of ice-sheets as assessed on the basis of post-AR5 methods[32]. The three scenarios represent the RCPs 2.6, 4.5 and 8.5, paired with a low, medium and high ice-sheet contribution, respectively, and generated using the general circulation model HadGEM2-ES[60] (Extended Data Table 2). The SLR scenarios are regionalized, therefore accounting for regional gravitational and rotational effects due to changes in ice mass distribution and steric variation[32]. Local relative SLR information is attained by combining the regionalized SLR projections with segment-specific vertical land movement based on a global model of glacial isostatic adjustment[61] and some additional 2 mm yr^{-1} of natural subsidence in large river deltas[62,63] (Extended Data Fig. 6). Meanwhile, human-induced subsidence, which may be of particular importance in large river deltas[64], is not considered for calculating regional relative SLR. However, a sensitivity analysis using delta-wide subsidence rates of 5 mm yr^{-1} showed only small deviation in modelled overall global wetland areas (Extended Data Table 4). Tectonic and neotectonic uplift and subsidence processes, other than glacial isostatic adjustment, are also not included owing to the lack of an appropriate global data set.

Tidal range data. To calculate the WAS (Extended Data Fig. 4) and compute the vertical wetland extent within each coastline segment, we use a newly developed global tidal range data set[65], representing the segment-specific tidal range (that is, the difference between mean low water and mean high water), mean high water neap (MHWN) and mean high water spring (MHWS) tidal levels. The new tidal dataset was generated using OTISmpi[66], a forward global tidal model, solving the nonlinear shallow water equations on a C-grid using a finite differences time stepping method (Supplementary Information).

Population density data. For each coastline segment, the coastal population within each elevation increment is computed by superimposing the SRTM digital elevation model[47] with the Global Rural-Urban Mapping Project (GRUMP) population data[67], being subject to national population growth according to SSP2 (IIASA)[20,68]. To determine the population density in the floodplain of the 1-in-100-year extreme water level, which is used as a proxy for the availability of accommodation space (Extended Data Fig. 4), we derive the hydrologically connected floodplain area for the 1-in-100-year extreme water level and the corresponding population affected by flooding[32]. We use the latest dataset on extreme water levels along the world's coastline, produced with a new global storm surge model hindcasting extreme water levels between 1979 and 2014[50]. Extreme water levels are reported for the return periods of 1, 10, 100 and 1,000 years and are derived from total water levels during storm surge events, thus including both tides and surges.

Sediment availability data. Local sediment availability is derived from MERIS satellite data, processed in the framework of the GlobColour project (http://globcolour.info). The data represent total suspended matter in the water column and have been developed, validated, and distributed by ACRI-ST, France[69]. We use the monthly averages from April 2002 to April 2012 that have a horizontal resolution of 1/24°. A long-term average is calculated for every pixel, and an average value of all pixels located within a 4 km buffer of each coastline segment is used to represent the local sediment availability (mg l^{-1}).

Conversion of terrestrial upland to coastal wetlands. With increasing sea levels, we allow coastal wetlands to migrate inland, a process that we understand as the establishment of wetland vegetation inland of its previous location, by raising the MHWS level along the coastal profile. Hence, former terrestrial upland areas are inundated and converted to coastal wetlands (Extended Data Fig. 5), based on elevation, where no human barriers are assumed to be present[36–39]. This modelling approach is supported by recent local-scale field studies for coastal salt marshes at the US east coast and in the Gulf of Mexico[70–74] and has previously been applied through various local-scale models, both for salt marshes and mangroves[75–79]. The establishment of coastal wetland vegetation in inundated upland areas is assumed to be associated with a response lag of five years, which is in line with evidence produced by recent wetland restoration studies[80–83]. However, the development of related wetland functions (such as biogeochemical functioning) may take longer[74,80].

For calculation of the converted upland areas, we assume the segment-specific wetland/non-wetland proportion to remain constant over time, whereby the non-wetland area within a coastline segment equals the total floodplain area (that is, the total interpolated area between MSL and MHWS) minus the reported wetland area. The conversion of uplands to wetlands is therefore calculated as the product of the wetland/non-wetland proportion and the total inundated upland area. However, conversion of terrestrial upland to coastal wetland is assumed to be zero where the coastal population density within the floodplain of the 1-in-100-year extreme water level exceeds the given thresholds (5, 20, 150 or 300 people km^{-2}), representing the existence of anthropogenic barriers to inland wetland migration. We thereby assume that coastal protection infrastructure is an important contributor to anthropogenic barriers for wetland inland migration[2,8,36–39] and is built where coastal communities are threatened by extreme water levels, such as a 1-in-100-year event[32,84].

Seaward loss of coastal wetlands. As sea level rises, not only the upper wetland boundary (MHWS) but also the lower wetland boundary (MSL) shifts position, potentially causing inundation of coastal wetlands beyond physiological tolerance. Therefore, we calculate an 'unconstrained seaward loss' that at first neglects the capacity of the wetland to adapt vertically to SLR by sediment accretion. Through sediment accretion, this unconstrained seaward loss may, however, be reduced or inhibited, given sufficient sediment availability within the coastline segment (Extended Data Fig. 4).

The WAS is a measure for the difference between the sediment needed for the coastal wetland to vertically accrete sediment as fast as SLR and the sediment available. It represents a sediment surplus if positive, and a sediment deficit if negative (Fig. 3). The amount of sediment needed for a coastal wetland to adapt to SLR has been studied previously[40], using an ensemble of five models for tidal marsh accretion. Linear relationships were presented between sediment availability and the maximum rate of relative SLR that a tidal marsh can survive, showing steeper slopes (higher resilience) for marshes in macrotidal environments compared to marshes in microtidal environments[40]. We directly use these linear relationships for our tidal marshes (including tidal salt and freshwater marshes), whereas we modify the model parameters for modelling mangrove forests during our calibration procedure (Supplementary Information). The local sediment availability, as derived from the GlobColour data, is assumed to represent the current levels of total suspended matter in the coastal zone and assumed to remain constant during the simulation period. To account for possible changes in future global sediment supply, a sensitivity analysis has been conducted with average sediment availability levels reduced and increased by 20% and 50% (Extended Data Table 3).

The WAS thus represents the ability of the coastal wetlands within a coastline segment to adapt to rising sea levels by sediment accretion. A positive WAS value indicates that sediment availability is sufficient to maintain the present wetland area, whereas a negative WAS value suggests that coastal wetlands are inundated and (partially) lost in response to SLR. The WAS is an integer value that ranges from −5 to +5, indicating a very high (−5) to very low (−1) sediment deficiency and a very low (+1) to very high (+5) sediment surplus, respectively. On the basis of the WAS, the unconstrained seaward loss ($\mathrm{SL_{unc}}$: km^2) is transformed into a constrained seaward loss ($\mathrm{SL_c}$: km^2), assuming a linear relationship between WAS and the proportion of inundated wetland actually being lost, but only if WAS is negative (equation (1)). No wetland loss is computed where WAS is positive or zero. With SLR, both the WAS and $\mathrm{SL_{unc}}$ change over time. Thus, $\mathrm{SL_c}$ is updated after every time step (t_i).

$$\mathrm{SL_c}(t_i) = (-1/5) \times \mathrm{WAS}(t_i) \times \mathrm{SL_{unc}}(t_i) \tag{1}$$

The calculation of WAS is based on the assumption that the critical rate of relative SLR ($\mathrm{RSLR_{crit}}$: mm yr^{-1}) depends on sediment availability (Sed: mg l^{-1}) and tidal range (TR), as suggested previously[40]. The previous modelling results[40] can be approximated using the following relationship (equation (2)):

$$\mathrm{RSLR_{crit}} = (m \times \mathrm{TR}^e) \times \mathrm{Sed} + i \tag{2}$$

in which ($m \times \mathrm{TR}^e$) represents the slope of a linear relationship between $\mathrm{RSLR_{crit}}$ and Sed. Model parameters e, i and m are calibrated separately for tidal marshes (including tidal salt and freshwater marshes, e_TF, i_TF and m_TF) and mangrove systems (e_Man, i_Man and m_Man). Parameters e_TF, i_TF and m_TF are directly derived from the model ensemble runs described previously[40], and e_Man, i_Man and m_Man are estimated by calibrating the model using the mangrove data presented previously[19] (Supplementary Information).

To estimate the sediment needed for a given SLR rate, $\mathrm{Sed_{crit}}$ (mg l^{-1}), we rewrite equation (2) as follows (equation (3)):

$$\mathrm{Sed_{crit}} = (\mathrm{RSLR}{-}i)/(m \times \mathrm{TR}^e) \tag{3}$$

in which RSLR (mm yr^{-1}) is the actual (time-dependent) local relative SLR rate. Knowing the current sediment availability (Sed) within each coastline segment (derived from the GlobColour data), we compare this value with the segment-specific $\mathrm{Sed_{crit}}$ and define WAS as the scaled and rounded difference between the available and needed sediment availability (equation (4)):

$$\mathrm{WAS} = \mathrm{round}\{[(\mathrm{Sed} - \mathrm{Sed_{crit}})/a] \times 5\} \tag{4}$$

in which a represents the sediment surplus (or deficit in case $\mathrm{Sedsup} < \mathrm{Sedsup_{crit}}$), which is considered as 'very high'. The determination of a is subject to model calibration (Supplementary Information). All WAS values greater (smaller) than 5 (−5) are transformed to WAS values of 5 (−5).

Model calibration. The model parameters m_TF, m_Man, e_TF, e_Man, i_TF, i_Man and a (equations (3) and (4)) are estimated using a stepwise calibration procedure as described in detail in the Supplementary Information. Model results are thereby compared to field measurements of vertical elevation growth for 39 marsh sites across US and European Atlantic shorelines[4], 18 marsh sites in North America, Europe and northeast Australia[3] and 26 mangrove sites across Pacific shorelines[19]. The calibrated model ($m_\mathrm{TF} = 3.42$, $m_\mathrm{Man} = 4.42$, $e_\mathrm{TF} = 0.915$, $e_\mathrm{Man} = 1.18$, $i_\mathrm{TF} = -1.5$, $i_\mathrm{Man} = 0$ and $a = 40$ mg l^{-1}) correctly predicts whether there is a sediment deficit, a sediment surplus or a balanced sediment budget for 78% of the coastline segments where field data are available (Extended Data Table 1).

Scenarios. The three SLR scenarios RCP 2.6, 4.5 and 8.5, accounting for the full range of available SLR scenarios[45], are combined with three human adaption scenarios. These are subject to population growth according to SSP2 (Extended Data Table 2), which is considered a middle-of-the-road scenario for population growth[68]. The three different human adaptation scenarios include a BAU scenario, and a moderate (NB 1) and an extreme (NB 2) nature-based adaptation scenario. They reflect differences in the potential of coastal wetlands to migrate inland until 2100 owing to potential differences in future coastal management strategies. In addition, four different physically and/or socio-economically unrealistic model configurations (Extended Data Table 2: hypothetical scenarios) are used during the sensitivity analysis to quantify the extent to which overall resilience is enabled/constrained by vertical and horizontal adaptability mechanisms, namely vertical sediment accretion and wetland inland migration.

Human adaptation scenarios. Inland/upward migration of coastal wetlands is often obstructed by the presence of anthropogenic infrastructure (for example, dikes, seawalls, cities, roads and railways)[18,37]. As there is no global dataset on coastal infrastructure, we approximate accommodation space through a population density threshold above which we assume that no accommodation space is available for coastal wetlands to migrate inland/upward. We thereby assume that coastal infrastructure is more likely to be present where population density is high[37,85], and that coastal protection structures are among the most important barriers for wetland inland migration[8]. By comparing a recent expert judgement on current coastal protection infrastructure, relying on population density and gross national income[86], with coastal population densities within the 1-in-100-year extreme water level floodplain, we find that currently, on a global average, coasts of >20 people km^{-2} are protected by some kind of coastal protection infrastructure (Supplementary Information). We consider this number as the upper boundary of current accommodation space. This is because it only includes coastal protection infrastructure and neglects other anthropogenic infrastructure that may act as barrier. As a lower boundary we choose a population density threshold of 5 people km^{-2} as this has previously been used to define (nearly) uninhabited land[87]. We therefore define the range of threshold population densities between 5 and 20 people km^{-2} as our BAU scenario (Fig. 1 and Extended Data Table 2).

In two nature-based adaptation scenarios (NB 1 and NB 2), we assume that coastal societies in rural areas retreat from the coast with SLR, removing coastal protection and other infrastructure that inhibit inland migration of coastal wetlands. We thereby assume that this is more likely to happen in sparsely populated areas as compared to densely populated areas[8,88–90]. For the first NB 1 scenario, we assume an upper boundary of 150 people km^{-2}, which corresponds to the Organisation for Economic Co-operation and Development (OECD) definition of urban areas[91]. In the second, more extreme NB 2 scenario, we use a threshold of 300 people km^{-2} as the upper boundary, as this corresponds to the European Commission's definition of urban areas[22] (Extended Data Table 2).

Hypothetical scenarios. The four hypothetical scenarios used for the sensitivity analysis include: (1) wetland migration only, characterized by the exclusion of bio-physical vertical accretion mechanisms and unlimited inland accommodation space; (2) sediment accretion only, characterized by the inclusion of bio-physical vertical accretion mechanisms, but assuming no inland accommodation space; (3) maximum resilience, which includes bio-physical accretion mechanisms and assumes an unlimited potential for inland migration; and (4) no resilience, in which neither bio-physical accretion nor inland migration are accounted for (Extended Data Table 2).

It should be noted that these hypothetical scenarios are unrealistic from a socio-economic or physical perspective, because no future coast will be neither completely defended nor completely undefended by dikes and seawalls and neither will sediment accretion be fully absent. But these hypothetical model runs are meant to demonstrate the relative contributions of the two mechanisms of wetland inland migration and sediment accretion to the overall wetland resilience to SLR.

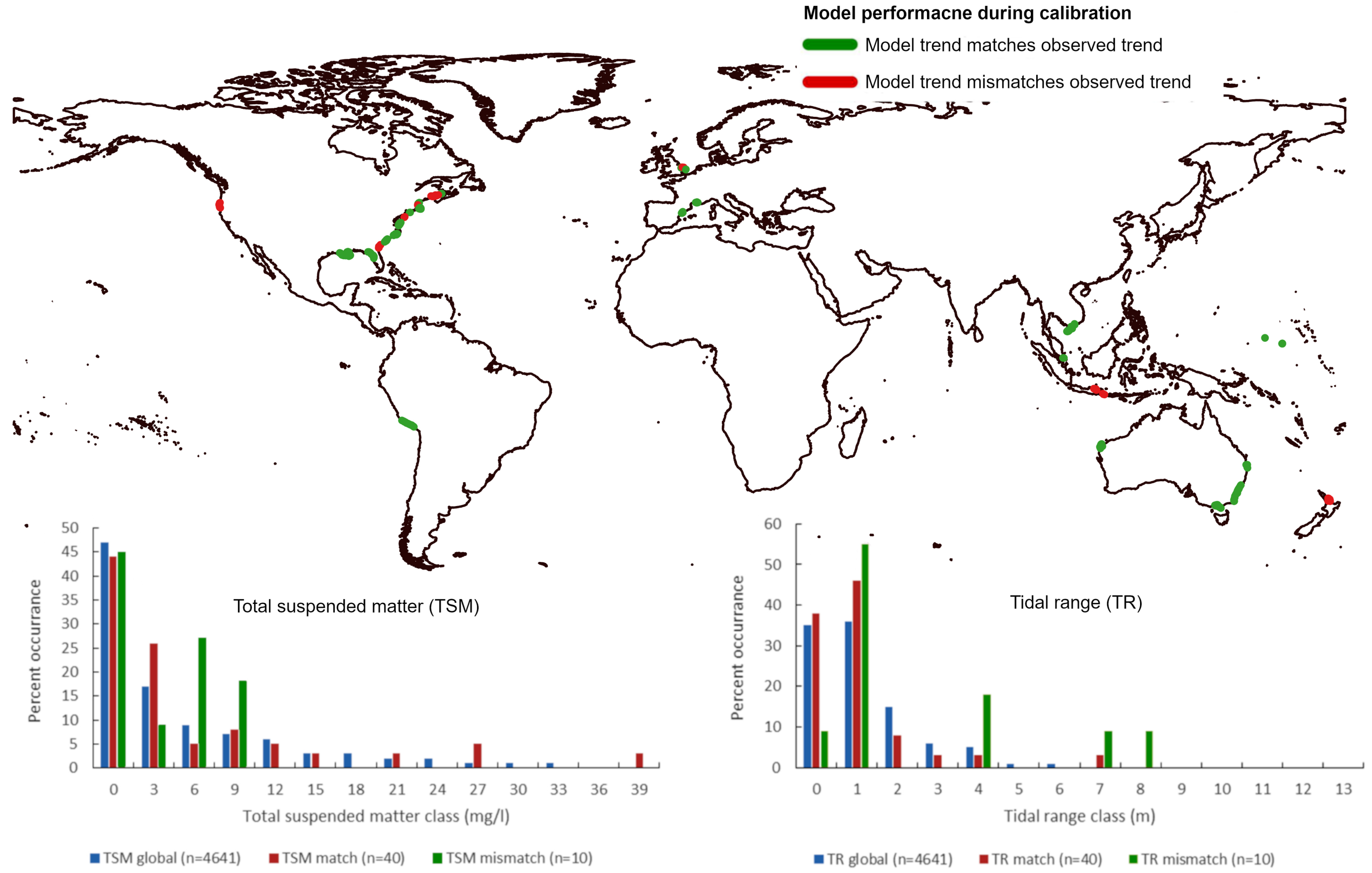
Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Code availability. The computer code that supports the findings of this study is available for non-commercial use (CC BY-NC-SA 4.0) from the GitLab repository 'global-coastal-wetland-model', https://gitlab.com/mark.schuerch/global-coastal-wetland-model.git. Registration on GitLab is required to access the repository.

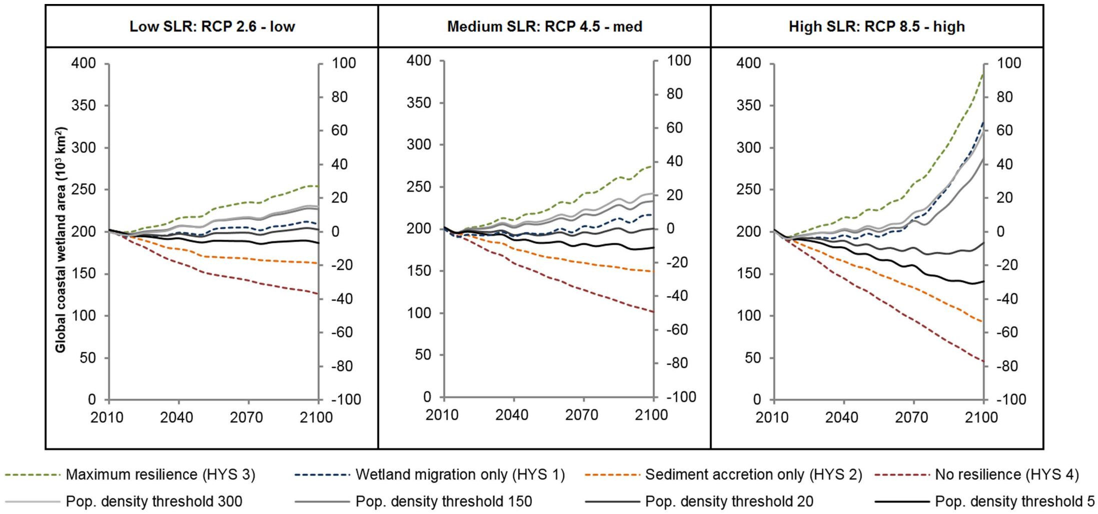
Data availability. The data that support the findings of this study are available from the corresponding author upon reasonable request. The Source Data for Fig. 1 and Extended Data Fig. 2 are provided.

31. Vafeidis, A. T. et al. A new global coastal database for impact and vulnerability analysis to sea-level rise. *J. Coastal Res.* **24**, 917–924 (2008).

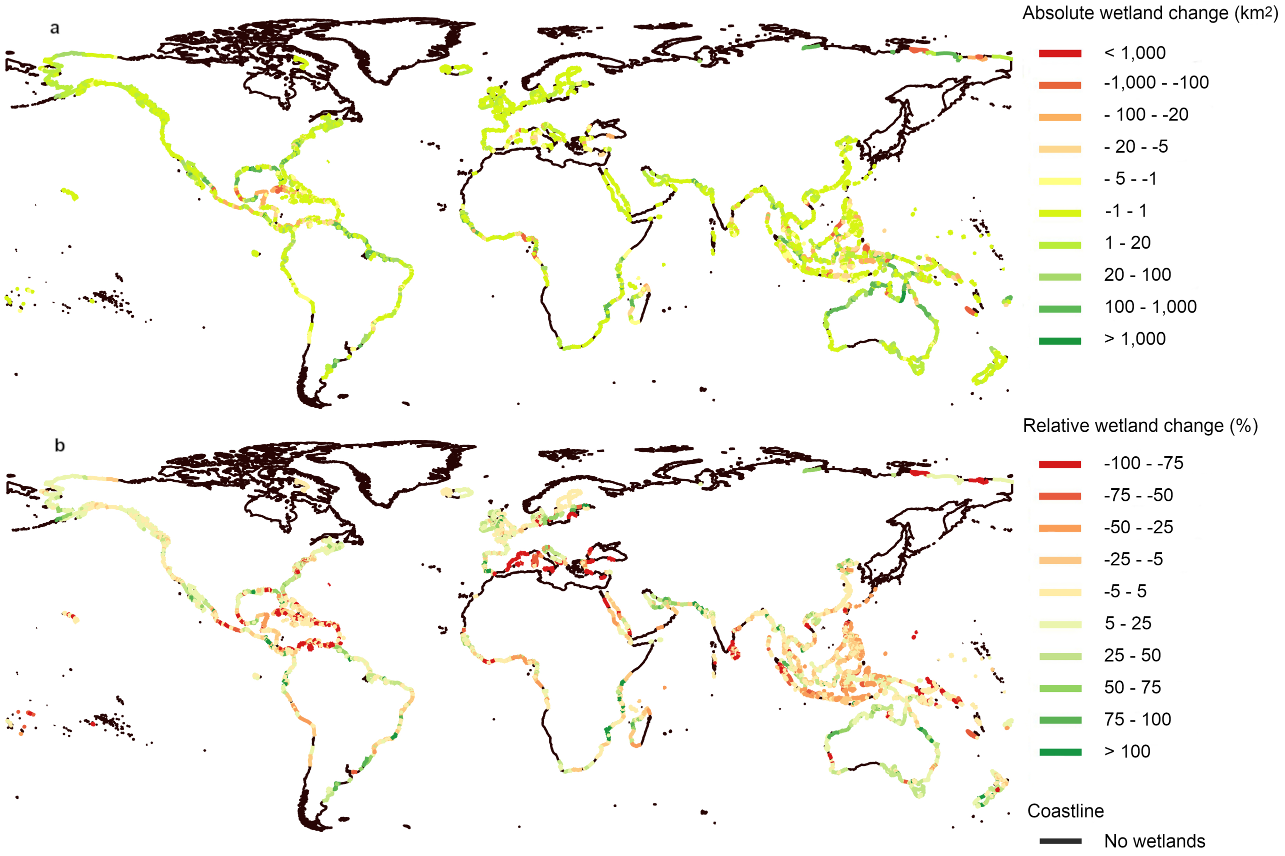
32. Hinkel, J. et al. Coastal flood damage and adaptation costs under 21st century sea-level rise. *Proc. Natl Acad. Sci. USA* **111**, 3292–3297 (2014).
33. Vafeidis, A. T. et al. Water-level attenuation in broad-scale assessments of exposure to coastal flooding: a sensitivity analysis. *Natural Hazards and Earth System Sciences Discussions* https://doi.org/10.5194/nhess-2017-199 (2017).
34. Giri, C. et al. Status and distribution of mangrove forests of the world using earth observation satellite data. *Glob. Ecol. Biogeogr*. **20**, 154–159 (2011).
35. McOwen, C. et al. A global map of saltmarshes. *Biodivers. Data J*. **5**, e11764 (2017).
36. Kirwan, M. L., Walters, D. C., Reay, W. G. & Carr, J. A. Sea level driven marsh expansion in a coupled model of marsh erosion and migration. *Geophys. Res. Lett*. **43**, 4366–4373 (2016).
37. Borchert, S. M., Osland, M. J., Enwright, N. M. & Griffith, K. T. Coastal wetland adaptation to sea level rise: quantifying potential for landward migration and coastal squeeze. *J. Appl. Ecol*. https://doi.org/10.1111/1365-2664.13169 (2018).
38. Gilman, E. L., Ellison, J., Duke, N. C. & Field, C. Threats to mangroves from climate change and adaptation options: a review. *Aquat. Bot*. **89**, 237–250 (2008).
39. Torio, D. D. & Chmura, G. L. Assessing Coastal Squeeze of Tidal *Wetlands. J. Coast. Res*. **29**, 1049–1061 (2013).
40. Kirwan, M. L. et al. Limits on the adaptability of coastal marshes to rising sea level. *Geophys. Res. Lett*. **37**, L23401 (2010).
41. D'Alpaos, A., Mudd, S. M. & Carniello, L. Dynamic response of marshes to perturbations in suspended sediment concentrations and rates of relative sea level rise. *J. Geophys. Res. Earth Surf*. **116**, F04020 (2011).
42. French, J. Tidal marsh sedimentation and resilience to environmental change: exploratory modelling of tidal, sea-level and sediment supply forcing in predominantly allochthonous systems. *Mar. Geol*. **235**, 119–136 (2006).
43. Kirwan, M. L. & Guntenspergen, G. R. Influence of tidal range on the stability of coastal marshland. *J. Geophys. Res. Earth Surf*. **115**, F02009 (2010).
44. Temmerman, S., Govers, G., Wartel, S. & Meire, P. Modelling estuarine variations in tidal marsh sedimentation: response to changing sea level and suspended sediment concentrations. *Mar. Geol*. **212**, 1–19 (2004).
45. Church, J. A. et al. in *Climate Change 2013: the Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (eds Stocker, T. F. et al.) 1137–1216 (Cambridge Univ. Press, Cambridge, 2013).
46. McFadden, L., Nicholls, R. J., Vafeidis, A. & Tol, R. S. J. Methodology for modeling coastal space for global assessment. *J. Coast. Res*. **23**, 911–920 (2007).
47. Jarvis, A., Reuter, H. I., Nelson, A. & Guevara, E. *Hole-Filled SRTM for the Globe Version 4*; http://srtm.csi.cgiar.org/ (2008).
48. Nicholls, R. J., Hoozemans, F. & Marchand, M. Increasing flood risk and wetland losses due to global sea-level rise: regional and global analyses. *Glob. Environ. Change* **9**, 69–87 (1999).
49. Nicholls, R. J. Coastal flooding and wetland loss in the 21st century: changes under the SRES climate and socio-economic scenarios. *Glob. Environ. Change* **14**, 69–86 (2004).
50. Muis, S., Verlaan, M., Winsemius, H. C., Aerts, J. C. J. H. & Ward, P. J. A global reanalysis of storm surges and extreme sea levels. *Nat. Commun*. **7**, 11969 (2016).
51. Titus, J. G. & Richman, C. Maps of lands vulnerable to sea level rise modeled elevations along the US Atlantic and Gulf coasts. *Clim. Res*. **18**, 205–228 (2001).
52. Titus, J. G. & Wang, J. in *Background Documents Supporting Climate Change Science Program Synthesis and Assessment Product 4.1 (EPA 430R07004)* (eds Titus, J. G. & Strange, E. M.) (United States Environmental Protection Agency (EPA), Washington DC, 2008).
53. Vafeidis, A. T., Nicholls, R. J., McFadden, L., Hinkel, J. & Grasshoff, P. S. Developing a global database for coastal vulnerability analysis: design issues and challenges. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci* **35**, 801–805 (2004).
54. Balke, T., Stock, M., Jensen, K., Bouma, T. J. & Kleyer, M. A global analysis of the seaward salt marsh extent: the importance of tidal range. *Wat. Resour. Res*. **52**, 3775–3786 (2016).
55. Ellison, J. in *Coastal Wetlands: an Integrated Ecosystem Approach* (eds Perillo, G. M. E. et al.) 565–591 (Elsevier, Amsterdam, 2009).
56. McIvor, A. L., Spencer, T., Möller, I. & M., S. *The Response of Mangrove Soil Surface Elevation to Sea Level Rise* (The Nature Conservancy and Wetlands International, 2013).
57. McKee, K. L. & Patrick, W. H. The relationship of smooth cordgrass (Spartina alterniflora) to tidal datums: a review. *Estuaries* **11**, 143–151 (1988).
58. Odum, W. E. Comparative ecology of tidal freshwater and salt marshes. *Annu. Rev. Ecol. Syst*. **19**, 147–176 (1988).
59. Gray, A. J., Marshall, D. F. & Raybould, A. F. A century of evolution in *Spartina anglica*. *Adv. Ecol. Res*. **21** 1–62 (1991).
60. Jones, C. D. et al. The HadGEM2-ES implementation of CMIP5 centennial simulations. *Geosci. Model Dev*. **4**, 543 (2011).
61. Peltier, W. Global glacial isostasy and the surface of the ice-age earth: the ice-5G (VM2) model and GRACE. *Annu. Rev. Earth Planet. Sci*. **32**, 111–149 (2004).
62. Meckel, T. A., Ten Brink, U. S. & Williams, S. J. Sediment compaction rates and subsidence in deltaic plains: numerical constraints and stratigraphic influences. *Basin Res*. **19**, 19–31 (2007).
63. Syvitski, J. P. M. Deltas at risk. *Sustain. Sci*. **3**, 23–32 (2008).
64. Ericson, J. P., Vörösmarty, C. J., Dingman, S. L., Ward, L. G. & Meybeck, M. Effective sea-level rise and deltas: causes of change and human dimension implications. *Global Planet. Change* **50**, 63–82 (2006).
65. Pickering, M. D. et al. The impact of future sea-level rise on the global tides. *Cont. Shelf Res*. **142**, 50–68 (2017).
66. Egbert, G. D., Ray, R. D. & Bills, B. G. Numerical modeling of the global semidiurnal tide in the present day and in the last glacial maximum. *J. Geophys. Res. Oceans* **109**, C03003 (2004).
67. Center for International Earth Science Information Network (CIESIN) Columbia University, International Food Policy Research Institute (IFPRI), The World Bank & Centro Internacional de Agricultura Tropical (CIAT). *Global Rural-Urban Mapping Project, Version 1 (GRUMPv1): Population Density Grid*; http:// doi.org/10.7927/H4R20Z93 (accessed 22 July 2016).
68. Fricko, O. et al. The marker quantification of the Shared Socioeconomic Pathway 2: a middle-of-the-road scenario for the 21st century. *Glob. Environ. Change* **42**, 251–267 (2017).
69. Barrot, G., Mangin, A. & Pinnock, S. *Global Ocean Colour for Carbon Cycle Research, Product User Guide* (ACRI-ST, Sophia-Antipolis, 2007).
70. Raabe, E. A. & Stumpf, R. P. Expansion of Tidal Marsh in Response to Sea-Level Rise: Gulf Coast of Florida, USA. *Estuaries Coasts* **39**, 145–157 (2016).
71. Schieder, N. W., Walters, D. C. & Kirwan, M. L. Massive upland to wetland conversion compensated for historical marsh loss in Chesapeake Bay, USA. *Estuaries Coasts* **41**, 940–951 (2018).
72. Smith, J. A. M. The role of *Phragmites australis* in mediating inland salt marsh migration in a mid-atlantic estuary. *PLoS One* **8**, e65091 (2013).
73. Langston, A. K., Kaplan, D. A. & Putz, F. E. A casualty of climate change? Loss of freshwater forest islands on Florida's Gulf Coast. *Glob. Change Biol*. **23**, 5383–5397 (2017).
74. Anisfeld, S. C., Cooper, K. R. & Kemp, A. C. Upslope development of a tidal marsh as a function of upland land use. *Glob. Change Biol*. **23**, 755–766 (2017).
75. Feagin, R. A., Martinez, M. L., Mendoza-Gonzalez, G. & Costanza, R. Salt marsh zonal migration and ecosystem service change in response to global sea level rise: a case study from an urban region. *Ecol. Soc*. **15**, 14 (2010).
76. Gilman, E., Ellison, J. & Coleman, R. Assessment of mangrove response to projected relative sea-level rise and recent historical reconstruction of shoreline position. *Environ. Monit. Assess*. **124**, 105–130 (2007).
77. Di Nitto, D. et al. Mangroves facing climate change: landward migration potential in response to projected scenarios of sea level rise. *Biogeosciences* **11**, 857–871 (2014).
78. Rogers, K., Saintilan, N. & Copeland, C. Managed retreat of saline coastal wetlands: challenges and opportunities identified from the Hunter River Estuary, Australia. *Estuaries Coasts* **37**, 67–78 (2014).
79. Stralberg, D. et al. Evaluating tidal marsh sustainability in the face of sea-level rise: a hybrid modeling approach applied to San Francisco Bay. *PLoS One* **6**, e27388 (2011).
80. Craft, C., Broome, S. & Campbell, C. Fifteen years of vegetation and soil development after brackish-water marsh creation. *Restor. Ecol*. **10**, 248–258 (2002).
81. Mossman, H. L., Brown, M. J. H., Davy, A. J. & Grant, A. Constraints on salt marsh development following managed coastal realignment: dispersal limitation or environmental tolerance? *Restor. Ecol*. **20**, 65–75 (2012).
82. Mossman, H. L., Davy, A. J. & Grant, A. Does managed coastal realignment create saltmarshes with 'equivalent biological characteristics' to natural reference sites? *J. Appl. Ecol*. **49**, 1446–1456 (2012).
83. Wolters, M., Garbutt, A., Bekker, R. M., Bakker, J. P. & Carey, P. D. Restoration of salt-marsh vegetation in relation to site suitability, species pool and dispersal traits. *J. Appl. Ecol*. **45**, 904–912 (2008).
84. Nicholls, R. J. et al. Stabilization of global temperature at 1.5°C and 2.0°C: implications for coastal areas. *Phil. Trans. A Math. Phys. Eng. Sci*. **376**, 20160448 (2018).
85. Song, J., Fu, X., Wang, R., Peng, Z.-R. & Gu, Z. Does planned retreat matter? Investigating land use change under the impacts of flooding induced by sea level rise. *Mitig. Adapt. Strategies Glob. Change* **23**, 703–733 (2017).
86. Sadoff, C. W. et al. *Securing Water, Sustaining Growth: Report of the GWP/OECD Task Force on Water Security and Sustainable Growth* (Univ. Oxford, Oxford, 2015).
87. Mittermeier, R. A. et al. Wilderness and biodiversity conservation. *Proc. Natl Acad. Sci. USA* **100**, 10309–10313 (2003).
88. Abel, N. et al. Sea level rise, coastal development and planned retreat: analytical framework, governance principles and an Australian case study. *Environ. Sci. Policy* **14**, 279–288 (2011).
89. Kousky, C. Managing shoreline retreat: a US perspective. *Clim. Change* **124**, 9–20 (2014).
90. Field, C. R., Dayer, A. A. & Elphick, C. S. Landowner behavior can determine the success of conservation strategies for ecosystem migration under sea-level rise. *Proc. Natl Acad. Sci. USA* **114**, 9134–9139 (2017).
91. The Organisation for Economic Co-operation and Development (OECD). *OECD Regional Typology* (OECD, Paris, 2011).


Model performacne during calibration
Model trend matches observed trend
Model trend mismatches observed trend
Total suspended matter (TSM)
Percent occurrance
Total suspended matter class (mg/l)
0 3 6 9 12 15 18 21 24 27 30 33 36 39
0 5 10 15 20 25 30 35 40 45 50
TSM global (n=4641)
TSM match (n=40)
TSM mismatch (n=10)
Tidal range (TR)
Percent occurrance
Tidal range class (m)
0 1 2 3 4 5 6 7 8 9 10 11 12 13
0 10 20 30 40 50 60
TR global (n=4641)
TR match (n=40)
TR mismatch (n=10)

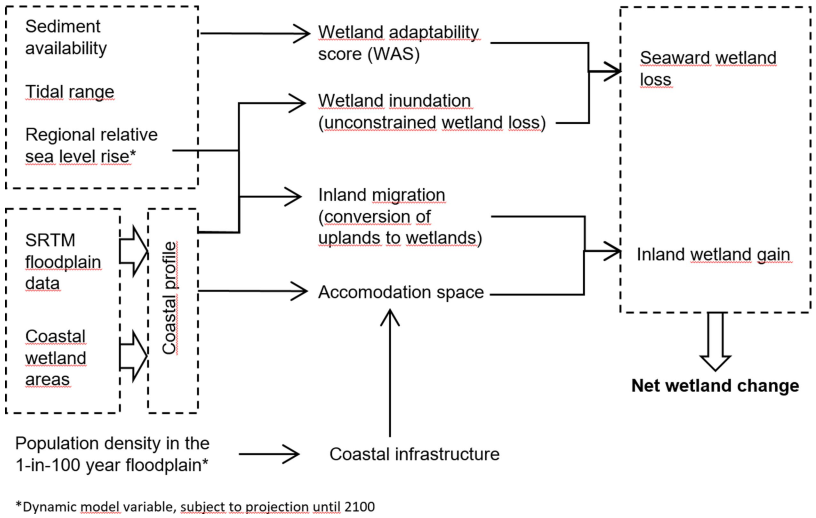
Extended Data Fig. 1 | Map of model performance during model calibration. Green lines indicate segments in which the modelled sediment balances match the observed trends in wetland elevation change relative to sea-level rise[3,4,19]. Red segments indicate model mismatches. The frequency distributions for total suspended matter (TSM) and tidal range (TR) display the distributions of both parameters in matching (green bars) and mismatching segments (red bars), and how they compare to the overall frequency distributions of both parameters (blue bars). The overall frequency distribution only includes coastline segments where coastal wetlands are present. The displayed coastline was generated during the DINAS-COAST FP5-EESD EU project (EVK2-CT-2000-00084).



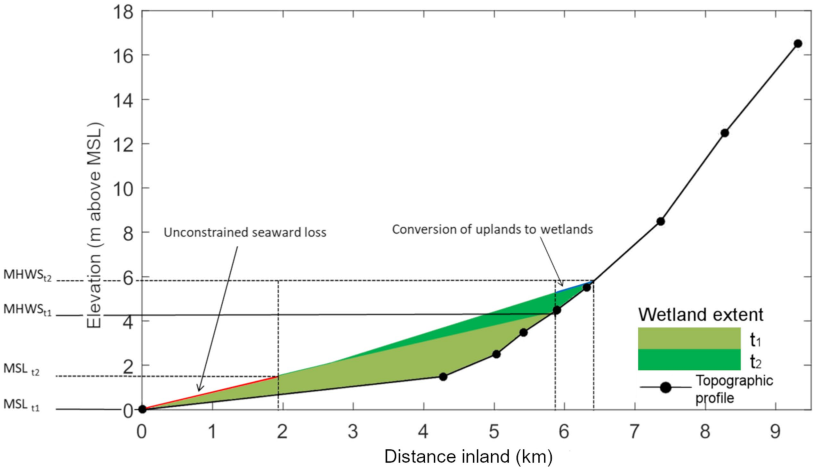
Extended Data Fig. 2 | Global change in coastal wetland area. Results for all three SLR scenarios (RCP 2.6, low; RCP 4.5, medium; RCP 8.5, high) and a total of eight different model configurations. These include the upper and lower boundaries of the BAU (5 and 20 people km^{-2}) and the upper boundaries of the NB 1 and NB 2 scenarios (150 and 300 people km^{-2}) as defined in Extended Data Table 2 (solid lines). The dashed lines represent the four hypothetical scenarios, as characterized in Extended Data Table 2: (i) wetland migration only; (ii) sediment accretion only; (iii) maximum resilience; and (iv) no resilience.


a
Absolute wetland change (km2)
< 1,000
-1,000 - -100
- 100 - -20
- 20 - -5
- 5 - -1
-1 - 1
1 - 20
20 - 100
100 - 1,000
> 1,000
b
Relative wetland change (%)
-100 - -75
-75 - -50
-50 - -25
-25 - -5
-5 - 5
5 - 25
25 - 50
50 - 75
75 - 100
> 100
Coastline
No wetlands

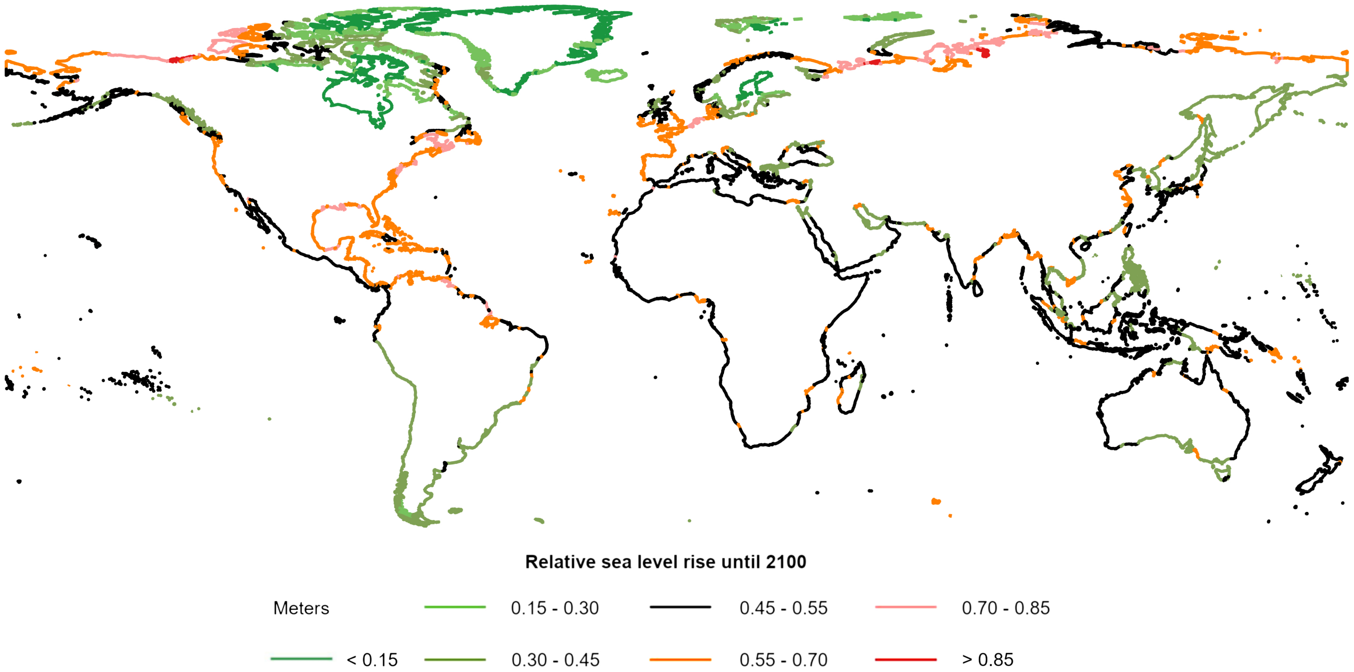
Extended Data Fig. 3 | Spatial distribution of coastal wetland change. **a**, **b**, Absolute (**a**) and relative (**b**) changes in coastal wetland areas are displayed for a medium SLR scenario (RCP 4.5)), assuming the possibility of wetland inland migration everywhere, but in urban areas with a population density more than 300 people km^{-2}. Population density is subject the population growth throughout the simulation period, following the Shared Socio-Economic Pathway SSP2[20,68]. The displayed coastline was generated during the DINAS-COAST FP5-EESD EU project (EVK2-CT-2000-00084).


Sediment availability
Tidal range
Regional relative sea level rise*
SRTM floodplain data
Coastal wetland areas
Coastal profile
Wetland adaptability score (WAS)
Wetland inundation (unconstrained wetland loss)
Inland migration (conversion of uplands to wetlands)
Accomodation space
Seaward wetland loss
Inland wetland gain
Net wetland change
Population density in the 1-in-100 year floodplain*
Coastal infrastructure
*Dynamic model variable, subject to projection until 2100

Extended Data Fig. 4 | Flow diagram representing the overall structure of the global coastal wetland model. Input parameters are shown on the left, output parameters are on the right. Net wetland change equals inland wetland gain minus seaward wetland loss.


Unconstrained seaward loss
Conversion of uplands to wetlands
MHWS$_{t2}$
MHWS$_{t1}$
MSL $_{t2}$
MSL $_{t1}$
Elevation (m above MSL)
0
2
4
6
8
10
12
14
16
18
Distance inland (km)
0
1
2
3
4
5
6
7
8
9
Wetland extent
t1
t2
Topographic profile

Extended Data Fig. 5 | Schematization of topographic profiles. The conversion of upland areas to coastal wetlands (if not inhibited by anthropogenic barriers) and the unconstrained seaward loss of coastal wetlands in response to sea-level rise is shown for an exemplary coastline segment (in western France). Inundation of terrestrial uplands follows the rising mean high water spring (MHWS) level between the time steps t_1 and t_2 (blue), whereas the unconstrained seaward loss follows the increase in mean sea level (MSL) when neglecting sediment accretion processes (red). To improve the clarity of the figure the actual MHWS level (2.54 m) and MSL rise are exaggerated.


Relative sea level rise until 2100
Meters
< 0.15
0.15 - 0.30
0.30 - 0.45
0.45 - 0.55
0.55 - 0.70
0.70 - 0.85
> 0.85

Extended Data Fig. 6 | Map of regionalized relative sea-level rise. Total relative sea-level rise (in m) for the medium SLR scenario (Extended Data Table 2) during the simulation period, including a delta subsidence rate of 2 mm yr^{-1} (2010–2100). Black coastlines indicate regions of relative sea-level rise similar to the global mean. The displayed coastline was generated during the DINAS-COAST FP5-EESD EU project (EVK2-CT-2000-00084).

Extended Data Table 1 | Performance of calibrated model when compared to field data

	Category (defined based on the field measurements)	Categorization of Field measurements	WAS	Model fit (number of segments)	Total occurrences (number of segments)
Tidal marshes	Total model fit	All data	All data	16	23
	“elevation deficit”	< -2 mm yr^{-1}	<-1	3	5
	“balanced”	-2 to 2 mm yr^{-1}	-1 to 1	10	10
	“elevation surplus”	> 2 mm yr^{-1}	>1	3	8
Mangrove systems	Total model fit	All data	All data	20	23
	“elevation deficit”	< -2 mm yr^{-1}	<-1	12	12
	“balanced”	-2 to 2 mm yr^{-1}	-1 to 1	6	6
	“elevation surplus”	> 2 mm yr^{-1}	>1	2	5

Summary of comparison between locally measured trends in surface elevation growth for tidal marshes[3,4] and mangrove systems[19] and modelled trends derived from the calculated WAS using $m_{\mathrm{TM}} = 3.42$, $m_{\mathrm{Man}} = 4.42$, $i_{\mathrm{TF}} = -1.5$, $i_{\mathrm{Man}} = 0$, $e_{\mathrm{TF}} = 0.915$, $e_{\mathrm{Man}} = 1.18$ and $a = 40$ mg l^{-1}. ‘Model fit’ represents the number of segments, where the calculated WAS corresponds with the measured category for vertical wetland growth.

Extended Data Table 2 | Characteristics of the scenarios

SLR scenario	Land-ice contribution	SLR until 2100 (cm)	Accommodation space scenario		Scenario	Population density threshold (people km^{-2})		Sediment accretion
						Lower boundary	Upper boundary	
Low: RCP 2.6 (5%)	low	29	Human adaptation scenarios	BAU[*]	BAU	5	20	Yes
				NB[†]	NB 1	20	150	
				NB[†]	NB 2	150	300	
			4 Hypothetical scenarios (HYS)		HYS 1	∞[‡]	∞[‡]	No
					HYS 2	0[§]	0[§]	Yes
					HYS 3	∞[‡]	∞[‡]	Yes
					HYS 4	0[§]	0[§]	No
Medium: RCP 4.5 (50%)	medium	50	Human adaptation scenarios	BAU[*]	BAU	5	20	Yes
				NB[†]	NB 1	20	150	
				NB[†]	NB 2	150	300	
			4 Hypothetical scenarios (HYS)		HYS 1	∞[‡]	∞[‡]	No
					HYS 2	0[§]	0[§]	Yes
					HYS 3	∞[‡]	∞[‡]	Yes
					HYS 4	0[§]	0[§]	No
High: RCP 8.5 (95%)	high	110	Human adaptation scenarios	BAU[*]	BAU	5	20	Yes
				NB[†]	NB 1	20	150	
				NB[†]	NB 2	150	300	
			4 Hypothetical scenarios (HYS)		HYS 1	∞[‡]	∞[‡]	No
					HYS 2	0[§]	0[§]	Yes
					HYS 3	∞[‡]	∞[‡]	Yes
					HYS 4	0[§]	0[§]	No

[*] BAU: Business-as-usual scenario

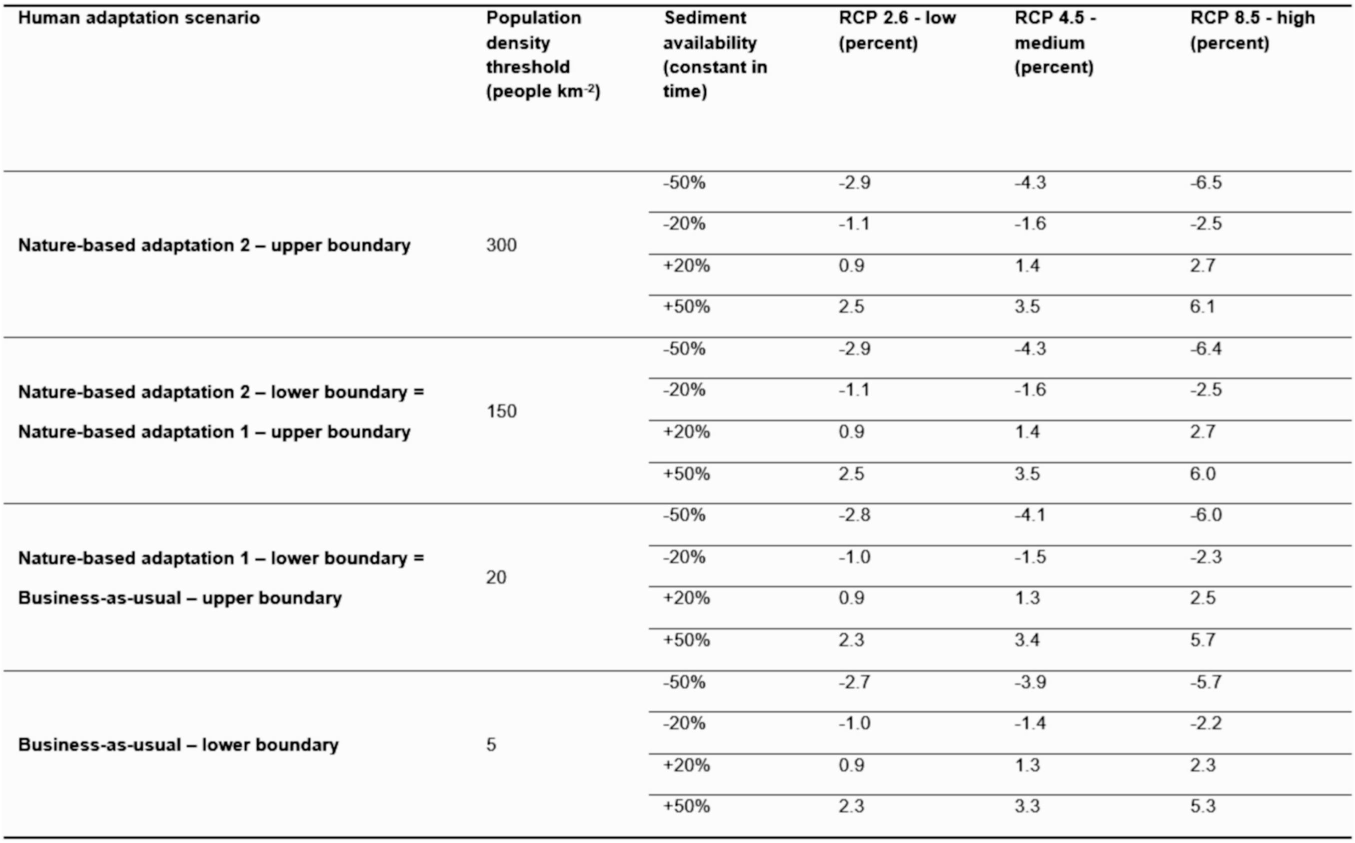
[†] NB: Nature-based adaptation scenarios

[‡] Population density threshold = ∞: Unlimited accommodation space

[§] Population density threshold = 0: No accommodation space

Three SLR scenarios (RCP 2.6, low; RCP 4.5, medium; RCP 8.5, high) were combined with three human adaptation scenarios (BAU; NB 1; and NB 2), accounting for varying degrees of accommodation space available for coastal wetlands, and four hypothetical scenarios (HYS 1: wetland migration only, HYS 2: sediment accretion only, HYS 3: maximum resilience, HYS 4: no resilience), used to quantify the contribution of vertical sediment accretion and horizontal inland migration to the overall resilience of coastal wetlands to global SLR (sensitivity analysis).

Extended Data Table 3 | Model sensitivity to variations in sediment availability



Human adaptation scenario	Population density threshold (people km^{-2})	Sediment availability (constant in time)	RCP 2.6 - low (percent)	RCP 4.5 - medium (percent)	RCP 8.5 - high (percent)
Nature-based adaptation 2 – upper boundary	300	-50%	-2.9	-4.3	-6.5
		-20%	-1.1	-1.6	-2.5
		+20%	0.9	1.4	2.7
		+50%	2.5	3.5	6.1
Nature-based adaptation 2 – lower boundary = **Nature-based adaptation 1 – upper boundary**	150	-50%	-2.9	-4.3	-6.4
		-20%	-1.1	-1.6	-2.5
		+20%	0.9	1.4	2.7
		+50%	2.5	3.5	6.0
Nature-based adaptation 1 – lower boundary = **Business-as-usual – upper boundary**	20	-50%	-2.8	-4.1	-6.0
		-20%	-1.0	-1.5	-2.3
		+20%	0.9	1.3	2.5
		+50%	2.3	3.4	5.7
Business-as-usual – lower boundary	5	-50%	-2.7	-3.9	-5.7
		-20%	-1.0	-1.4	-2.2
		+20%	0.9	1.3	2.3
		+50%	2.3	3.3	5.3

Percentage deviations in total global wetland area by 2100 from simulations with current-day sediment availability for all four population density thresholds (Extended Data Table 2) and reductions/increases of the constant sediment supply by 50% and 20%.

Extended Data Table 4 | Model sensitivity to variations in natural and human-induced delta subsidence

Model setup	Delta subsidence (mm yr^{-1})	RCP 2.6 - 5% (percent)	RCP 4.5 - 50% (percent)	RCP 8.5 - 95% (percent)
Pop. density threshold 300	0	14.0	18.7	55.8
Pop. density threshold 150		11.4	14.8	39.5
Pop. density threshold 20		0.1	-1.0	-10.2
Pop. density threshold 5		-6.7	-10.6	-31.3
Pop. density threshold 300	2	15.2	19.8	59.7
Pop. density threshold 150		12.3	15.3	42.4
Pop. density threshold 20		0.2	-0.8	-7.6
Pop. density threshold 5		-7.6	-11.9	-30.3
Pop. density threshold 300	5	16.7	21.5	62.7
Pop. density threshold 150		12.8	15.5	44.1
Pop. density threshold 20		0.5	1.2	-6.3
Pop. density threshold 5		-9.6	-14.6	-31.0

Percentage gain (positive) and loss (negative) of total global wetland area by 2100 from simulations for all four population density thresholds (Extended Data Table 2) and three different rates for uniform delta subsidence for all 117 deltas listed in the DIVA database[31].

LETTER

https://doi.org/10.1038/s41586-018-0445-z

Stepwise and independent origins of roots among land plants

Alexander J. Hetherington[1] & Liam Dolan[1]*

Roots are one of the three fundamental organ systems of vascular plants[1], and have roles in anchorage, symbiosis, and nutrient and water uptake[2–4]. However, the fragmentary nature of the fossil record obscures the origins of roots and makes it difficult to identify when the sole defining characteristic of extant roots—the presence of self-renewing structures called root meristems that are covered by a root cap at their apex[1–9]—evolved. Here we report the discovery of what are—to our knowledge—the oldest meristems of rooting axes, found in the earliest-preserved terrestrial ecosystem[10] (the 407-million-year-old Rhynie chert). These meristems, which belonged to the lycopsid *Asteroxylon mackiei*[11–14], lacked root caps and instead developed a continuous epidermis over the surface of the meristem. The rooting axes and meristems of *A. mackiei* are unique among vascular plants. These data support the hypothesis that roots, as defined in extant vascular plants by the presence of a root cap[7], were a late innovation in the vascular lineage. Roots therefore acquired traits in a stepwise fashion. The relatively late origin in lycophytes of roots with caps is consistent with the hypothesis that roots evolved multiple times[2] rather than having a single origin[1], and the extensive similarities between lycophyte and euphyllophyte roots[15–18] therefore represent examples of convergent evolution. The key phylogenetic position of *A. mackiei*—with its transitional rooting organ—between early diverging land plants that lacked roots and derived plants that developed roots demonstrates how roots were 'assembled' during the course of plant evolution.

The body plan of *A. mackiei* comprised three types of axes; leafy shoot axes, rooting axes and the transition region between these axes[11–13,19,20]. To identify meristems at the apices of *A. mackiei* rooting axes, we visually inspected 641 thin sections prepared from the Rhynie chert. We discovered the apices of five rooting axes among three thin sections (Fig. 1a–e). These five apices were assigned to *A. mackiei* on the basis of two pieces of evidence. First, each of the five apices was discovered on thin sections in which *A. mackiei* was the only plant species present (Extended Data Fig. 1). Second, the morphology of the G-type tracheids in two of the rooting axes is diagnostic of *A. mackiei*[21,22] (white arrowheads, Fig. 1f, g). These data indicate that the five apices are the tips of axes of *A. mackiei*.

Three characteristics enabled us to identify the five apices as the rooting axes of *A. mackiei*. First, *A. mackiei* rooting axes lacked leaves and stomata[11–13,19] (Fig. 1a). Stomata, leaf primordia and leaves were not present on either apical or subapical surfaces on any of the apices (Fig. 1a–e). Second, *A. mackiei* rooting axes grew in the direction of the gravity vector[11–13,19]. The positively gravitropic growth of two apices (Fig. 1a, b) was inferred from their orientation relative to sediment layers in both the growth substrate and a geopetally infilled void preserved in the thin section (Extended Data Fig. 2). The two axes that terminated in apices (Fig. 1a, b, Extended Data Fig. 2) were orientated perpendicular to the sediment layers; this indicated that the axes grew gravitropically into the sediment. Third, *A. mackiei* rooting axes were isotomously branching systems[11–13,19]. The organization of the pair of apices on the rooting axis in Fig. 1b indicates that they had recently branched isotomously. Taken together, the lack of leaves and stomata, the positively gravitropic growth and isotomous branching indicate that we identified five apices of *A. mackiei* rooting axes.

Having verified the identity of the *A. mackiei* apices, we next characterized the organization of tissues in these meristems and compared their organization with those of the roots of extant vascular plants. Three fundamental tissues were present in the apices of *A. mackiei* rooting axes (Fig. 1). Vascular tissue developed at the centre of the axis and was surrounded by layers of ground tissue; these two tissues were covered by a distinct epidermis that comprised cells that divided anticlinally (Fig. 1, Extended Data Fig. 3). None of the epidermal cells developed hair cells that resemble rhizoids or root hairs. The region in which these tissues converged on the tip constituted the promeristem[23]. The promeristem was multicellular, similar to the root meristems of many extant lycopsids[17,24] and the shoot meristem of *A. mackiei*[25]. In all five cases the promeristem contained the apical-most cells of the axis; there was no tissue beyond the apical region of the promeristem, which is where a root cap would be located in roots of extant vascular plants[17,23,24,26]. The roots of extant vascular plants comprise four tissues—vascular, ground, epidermis and root cap—that are derived from the promeristem[17,23,24,26]. The vascular tissue, ground tissues and epidermis differentiate basally, as in *A. mackiei*. However, a root cap also develops apically and laterally in extant vascular plants[17,23,24,26]. The discovery that five apices of the rooting axes of *A. mackiei* lacked root caps suggests that the rooting axes of early diverging lycopsids did not develop root caps.

Given that all extant plant species develop root caps at their root apices[7,23,24,26], we sought explanations that might account for the absence of the root cap in these fossils. We first considered that taphonomic processes could have led to the selective preservation of vascular, ground and dermal tissues without the preservation of root caps. Three of the apices (Fig. 1c–e) were preserved growing through a thin layer of degraded organic material called mulm[27], which coats the apex and flanks. It is formally possible that the mulm around the apex of the rooting structure could represent a decayed root cap. However, mulm also coats the external surfaces of many plants preserved in the Rhynie chert[27], including leafy shoots of *A. mackiei* preserved on the same thin sections as the three apices (Fig. 1c–e, Extended Data Fig. 4). The preservation of mulm around leafy shoots indicates that the presence of this substance at the apices of rooting axes does not represent the remains of root caps. Furthermore, given that all other tissues in the apices were well-preserved, it is likely that root caps would have been preserved if they were present (Fig. 1). Finally, root caps were readily preserved in permineralized plant deposits in the Carboniferous and Permian periods, which demonstrates that root caps can be fossilized[28,29]. We therefore rule out the possibility that root caps formed but were not preserved.

An alternative explanation for the lack of root caps in these apices is that the root cap may not have been present on apices that had stopped growing. We therefore established whether the apices were fossilized during active growth or after growth had ceased. During active growth, the root meristems of extant plants comprise large numbers of relatively small, dividing cells, and cells increase in size with distance from the apex as they elongate and differentiate[23,28,30]. In three

---

[1]Department of Plant Sciences, University of Oxford, Oxford, UK. *e-mail: liam.dolan@plants.ox.ac.uk

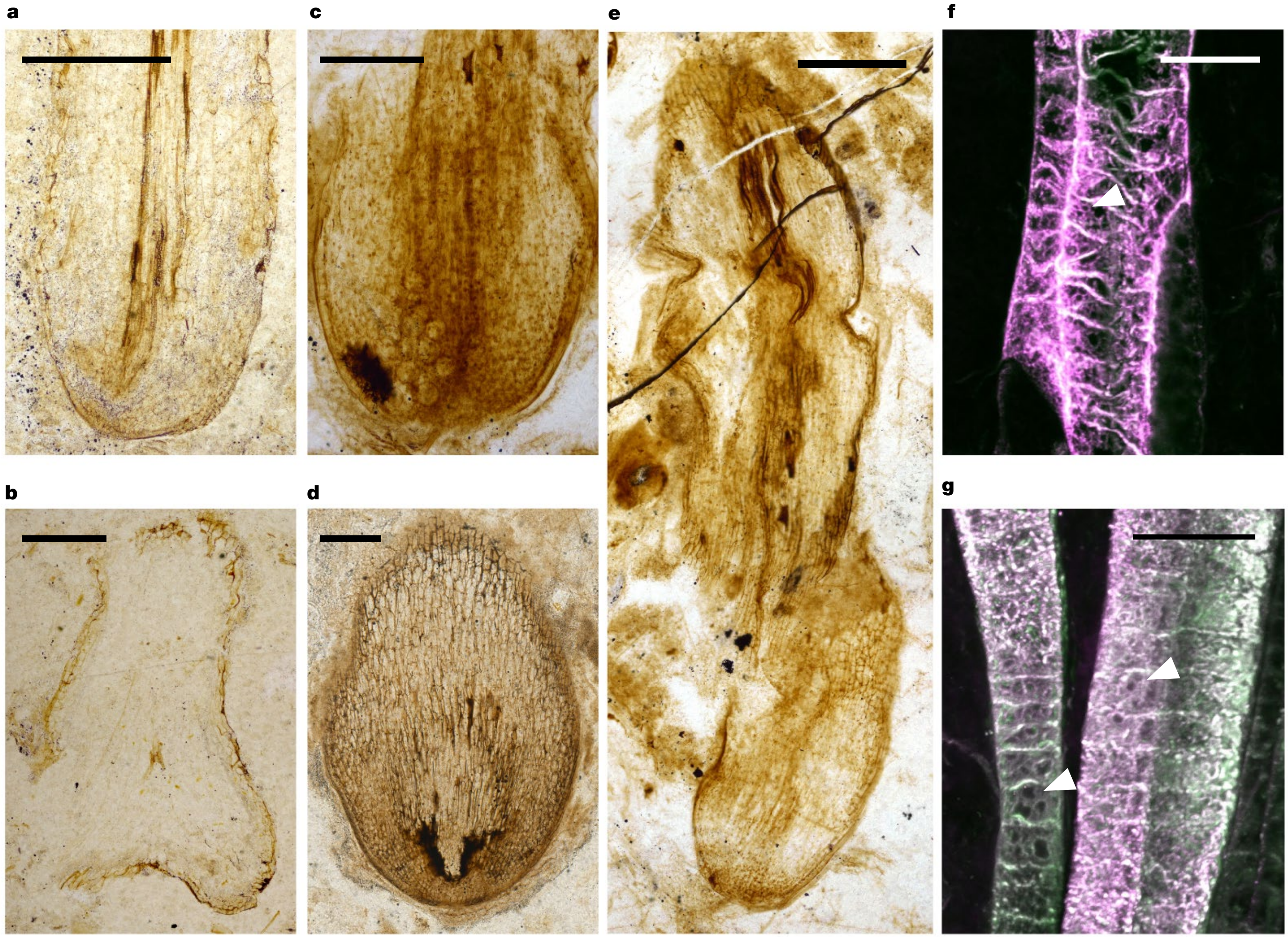


Fig. 1 | Five apices of the rooting axes of *A. mackiei* that lack root caps. **a**–**e**, The tips of five rooting axes of *A. mackiei* that terminate in apices. All apices are covered by a continuous layer of tissue and lack the tapered root caps that are characteristic of the roots of extant vascular plants. **f**, **g**, Confocal image of the tracheids in the axes shown in **a** (**f**) and **e** (**g**). The sheet perforations between thickening (arrowheads) are a characteristic of G-type tracheids that is diagnostic of *A. mackiei*. **a**, **b**, **f**, Specimen accession code: Natural History Museum, London (NHMUK) V.15642. **c**, **e**, **g**, Specimen accession code: The Hunterian, University of Glasgow (GLAHM) Kid 3080. **d**, Specimen accession code: Oxford University Herbaria (OXF) 108. Scale bars, 500 μm (**a**, **b**, **e**), 250 μm (**c**, **d**), 20 μm (**f**, **g**).

of the five *A. mackiei* apices, large numbers of small cells were located at the apex and cell size gradually increased with distance from the tip (Fig. 1c–e, Extended Data Fig. 3). This cellular organization indicated that these three apices were actively growing when fossilized. In the two remaining apices, the differentiation of vascular tissue near the tip of the apex[28,30] indicated that neither apex was active at the time of preservation (Fig. 1a, b, Extended Data Fig. 3). The absence of root caps from the three meristems that were actively growing when fossilized is consistent with the hypothesis that root caps did not develop on *A. mackiei* rooting axes.

If root caps developed in *A. mackiei*, there would be evidence in the preserved promeristems of the cell divisions from which the root caps developed. We characterized the cellular organization of the two well-preserved promeristems that were active when fossilized (Fig. 1c, d). We determined the orientations of cell walls at the apex to test whether these were consistent with the development of a root cap[23]. If a root cap was formed, periclinal divisions (in which the new wall is parallel to the surface) would occur near the apex of the promeristem[23]. Figure 2a, b shows an oblique section through the meristem that preserves the apical promeristem. There are no periclinal divisions in the outer layer of the apical promeristem (Fig. 2a, b). Instead, cells in the outermost layer of the apical region divided anticlinally (new cell walls are oriented perpendicular to the surface). This indicates that the cell-division pattern of the apical region of the meristem was inconsistent with the development of a root cap, and was instead consistent with the development of a continuous epidermis over the surface of the meristem. If a lateral root cap formed, it would develop from a combination of both anticlinal and periclinal cell divisions and produce layers of cells that taper with distance from the apex, as cells are sloughed off[23]. Figure 2c, d shows a longitudinal section through the meristem, in which the outer layer of the apex is preserved. Only anticlinal cell divisions occur in the outer layer and there is no evidence of periclinal divisions. The cell-division pattern of these meristems is inconsistent with the development of a lateral root cap, and is instead consistent with the development of a continuous epidermal surface. Together, the cellular organization of two promeristems enabled us to rule out the development of a root cap in *A. mackiei*; we predict that *A. mackiei* meristems were covered by a continuous epidermis.

To test the hypothesis that a continuous epidermal surface covered the meristems of rooting axes of *A. mackiei*, we reconstructed a three-dimensional model of the surface of the meristem from a *z*-stack of images captured on a confocal microscope (Fig. 3, Supplementary Videos 1, 2). A continuous and smooth layer of epidermis covered the meristem and there was no evidence of tapering or cells sloughing off. We conclude that meristems of the rooting axes of *A. mackiei* developed a continuous epidermis that covered the apex.

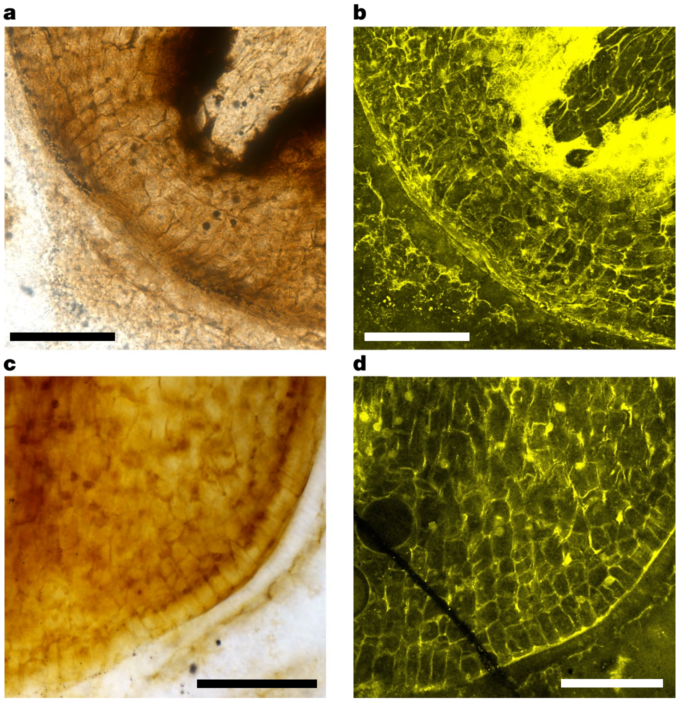


Fig. 2 | Cell division patterns in meristems of the rooting axes of *A. mackiei* are inconsistent with the formation of root caps. **a**–**d**, Transmitted light (**a**, **c**) and confocal laser (**b**, **d**) microscopy images showing the cellular organization of two meristems of the rooting axes of *A. mackiei*. **a**, **b**, The absence of periclinal cell divisions in the apical promeristem is inconsistent with the formation of a apical root cap. **c**, **d**, The absence of periclinal cell divisions on the flanks of the meristem is inconsistent with the formation of a lateral root cap. Anticlinal cell divisions in both the promeristem (**a**, **b**) and lateral flanks (**c**, **d**) are consistent with the meristem being covered by a continuous epidermis. **a**, **b**, OXF 108 (Fig. 1d). **c**, **d**, GLAHM Kid 3080 (Fig. 1c). Scale bars, 100 μm.

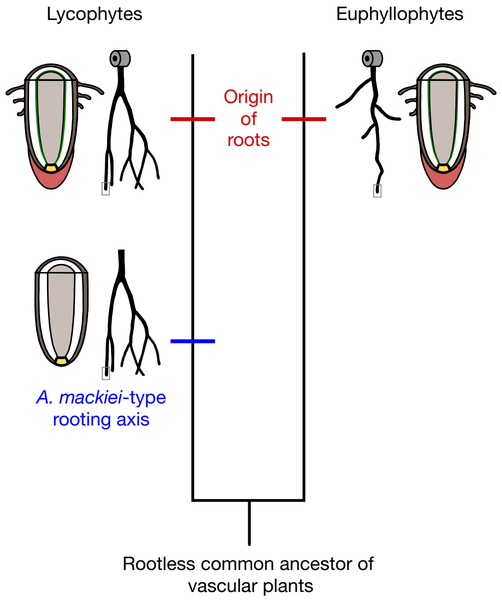

Lycophytes
Euphyllophytes
Origin of roots
A. mackiei-type rooting axis
Rootless common ancestor of vascular plants

Fig. 4 | Stepwise and independent origins of roots among vascular plants. The roots of extant lycophytes evolved in a stepwise manner. There was a stage—represented by *A. mackiei*— that was characterized by the presence of radially symmetric, positively gravitropic, isotomously branching axes. Subsequently, a root cap, root hairs, endogenous development and an endodermis all evolved. This sequence of events is consistent with the hypothesis that the common ancestor of all extant vascular plants was rootless, and roots with caps had at least two independent origins among lycophytes and euphyllophytes. The extensive similarities between roots of extant lycophytes and euphyllophytes therefore represent examples of convergent evolution.

The cellular organization of the five apices demonstrates that the rooting axes of *A. mackiei* developed from a previously unknown type of meristem, which lacked both root caps and root hairs. We conclude that the evolution of rooting axes in lycopsids occurred in a stepwise manner. There was a stage—represented by *A. mackiei*—that was characterized by the presence of radially symmetric, positively gravitropic, isotomously branching axes with vascular tissue organization that is distinct from that in the leafy shoot, as well as by a lack of leaves and stomata[11–13,19] (Fig. 4). Subsequently, a root cap, root hairs, endogenous development and an endodermis all evolved[16–18,24] (Fig. 4). This sequence of events is consistent with the hypothesis that the common ancestor of all extant vascular plants was rootless and that roots with caps had at least two separate origins, among lycophytes and euphyllophytes, respectively[2] (Fig. 4). The similarities in root anatomy (the presence of a root cap, root hairs, an endodermis and quiescent centre[16–18,24]), development (endogenous origin)[18] and gene expression[15] shared by extant lycophytes and euphyllophytes therefore represent examples of convergent evolution. The simultaneous presence of rooting axes and leaves on shoots in *A. mackiei* suggests that the co-evolution of roots and leaf-bearing shoots may have contributed to the evolution of an efficient transpiration stream among early lycopsids.

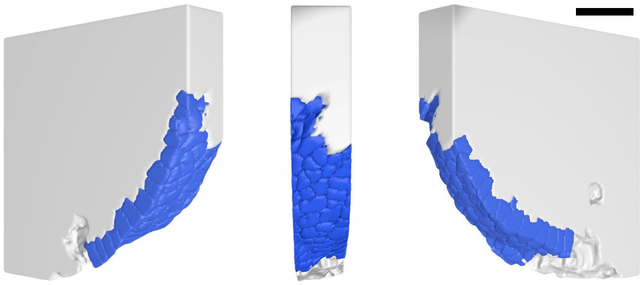


Fig. 3 | Meristems of the rooting axes of *A. mackiei* were covered by a continuous layer of epidermis and lacked a root cap. Three-dimensional model—from a *z*-stack of images captured on a confocal microscope—of the meristem of a rooting axis of *A. mackiei*, with the epidermal layer highlighted in blue. The cells in the epidermis form a smooth continuous surface over the meristem. Model based on images of GLAHM Kid 3080 (Figs. 1c, 2c, d). Scale bar, 50 μm.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0445-z.

Received: 18 April 2018; Accepted: 10 July 2018;
Published online 22 August 2018.

1. Schneider, H., Pryer, K. M., Cranfill, R., Smith, A. R. & Wolf, P. G. in *Developmental Genetics and Plant Evolution* (eds Cronk, Q. C. B. et al.) 330–364 (Taylor & Francis, London, 2002).
2. Raven, J. A. & Edwards, D. Roots: evolutionary origins and biogeochemical significance. *J. Exp. Bot.* **52**, 381–401 (2001).
3. Kenrick, P. & Strullu-Derrien, C. The origin and early evolution of roots. *Plant Physiol.* **166**, 570–580 (2014).
4. Kenrick, P. in *Plant Roots: The Hidden Half* (eds Eshel, A. & Beeckamn, T.) 1–14 (Taylor & Francis, Boca Raton, 2013).
5. Gensel, P. G., Kotyk, M. E. & Brasinger, J. F. in *Plants Invade The Land: Evolutionary and Environmental Perspectives* (eds Gensel, P. G. & Edwards, D.) 83–102 (Columbia Univ. Press, New York, 2001).
6. Matsunaga, K. K. S. & Tomescu, A. M. F. Root evolution at the base of the lycophyte clade: insights from an Early Devonian lycophyte. *Ann. Bot.* **117**, 585–598 (2016).
7. Sachs, J. *Text-book of Botany, Morphological and Physiological* (translated and annotated by A. W. Bennett and W. T. Thiselton Dyer) (Clarendon, Oxford, 1875).
8. Hao, S., Xue, J., Guo, D. & Wang, D. Earliest rooting system and root:shoot ratio from a new *Zosterophyllum* plant. *New Phytol.* **185**, 217–225 (2010).

9. Matsunaga, K. K. S. & Tomescu, A. M. F. An organismal concept for *Sengelia radicans* gen. et sp. nov. – morphology and natural history of an Early Devonian lycophyte. *Ann. Bot.* **119**, 1097–1113 (2017).
10. Edwards, D., Kenrick, P. & Dolan, L. History and contemporary significance of the Rhynie cherts—our earliest preserved terrestrial ecosystem. *Phil. Trans. R. Soc. B* **373**, 20160489 (2018).
11. Edwards, D. Embryophytic sporophytes in the Rhynie and Windyfield cherts. *Trans. R. Soc. Edinb. Earth Sci*. **94**, 397–410 (2003).
12. Kidston, R. & Lang, W. H. On Old Red Sandstone plants showing structure, from the Rhynie Chert Bed, Aberdeenshire. Part III. *Asteroxylon mackiei*, Kidston and Lang. *Trans. R. Soc. Edinb. Earth Sci*. **52**, 643–680 (1920).
13. Kidston, R. & Lang, W. H. On Old Red Sandstone plants showing structure, from the Rhynie Chert Bed, Aberdeenshire. Part IV. Restorations of the vascular cryptogams, and discussion of their bearing on the general morphology of the Pteridophyta and the origin of the organisation of land plants. *Trans. R. Soc. Edinb. Earth Sci*. **52**, 831–854 (1921).
14. Kenrick, P. & Crane, P. R. *The Origin and Early Diversification of Land Plants: A Cladistic Study (Smithsonian Series in Comparative Evolutionary Biology)* (Smithsonian Institute, Washington DC, 1997).
15. Huang, L. & Schiefelbein, J. Conserved gene expression programs in developing roots from diverse plants. *Plant Cell* **27**, 2119–2132 (2015).
16. Hetherington, A. J. & Dolan, L. The evolution of lycopsid rooting structures: conservatism and disparity. *New Phytol*. **215**, 538–544 (2017).
17. Fujinami, R. et al. Root apical meristem diversity in extant lycophytes and implications for root origins. *New Phytol*. **215**, 1210–1220 (2017).
18. Foster, A. S. & Gifford, E. M. *Comparative Morphology of Vascular Plants* (W. H. Freeman, San Francisco, 1959).
19. Bhutta, A. A. *Studies on the Flora of the Rhynie Chert.* PhD thesis, Univ. Wales, Cardiff (1969).
20. Kerp, H. Organs and tissues of Rhynie chert plants. *Phil. Trans. R. Soc. B* **373**, 20160495 (2018).
21. Kenrick, P. & Crane, P. R. Water-conducting cells in early fossil land plants: implications for the early evolution of tracheophytes. *Bot. Gaz.* **152**, 335–356 (1991).
22. Strullu-Derrien, C., Wawrzyniak, Z., Goral, T. & Kenrick, P. Fungal colonization of the rooting system of the early land plant *Asteroxylon mackiei* from the 407-Myr-old Rhynie Chert (Scotland, UK). *Bot. J. Linn. Soc*. **179**, 201–213 (2015).
23. Clowes, F. A. L. *Apical Meristems* (Blackwell, Oxford, 1961).
24. Guttenberg, H. V. *Histogenese der Pteridophyten. Handbuch der Pflanzenanatomie* Vol. VII.2 (Gebrüder Borntraeger, Berlin, 1966).
25. Hueber, F. M. Thoughts on the early lycopsids and zosterophylls. *Ann. Mo. Bot. Gard*. **79**, 474–499 (1992).
26. Heimsch, C. & Seago, J. L. Jr. Organization of the root apical meristem in angiosperms. *Am. J. Bot*. **95**, 1–21 (2008).
27. Kelman, R., Feist, M., Trewin, N. H. & Hass, H. Charophyte algae from the Rhynie chert. *Trans. R. Soc. Edinb. Earth Sci*. **94**, 445–455 (2003).
28. Hetherington, A. J., Dubrovsky, J. G. & Dolan, L. Unique cellular organization in the oldest root meristem. *Curr. Biol*. **26**, 1629–1633 (2016).
29. Strullu-Derrien, C., McLoughlin, S., Philippe, M., Mørk, A. & Strullu, D. G. Arthropod interactions with bennettitalean roots in a Triassic permineralized peat from Hopen, Svalbard Archipelago (Arctic). *Palaeogeogr. Palaeoclimatol. Palaeoecol*. **348–349**, 45–58 (2012).
30. Shishkova, S., Rost, T. L. & Dubrovsky, J. G. Determinate root growth and meristem maintenance in angiosperms. *Ann. Bot*. **101**, 319–340 (2008).

Acknowledgements A.J.H. was funded by the George Grosvenor Freeman Fellowship by Examination in Sciences, Magdalen College (Oxford). L.D. was funded by a European Research Council Advanced Grant (EVO500, contract 250284) and a European Commission Framework 7 Initial Training Network (PLANTORIGINS, project identifier 238640). We are grateful to the Oxford University Herbaria, the University of Manchester, Manchester Museum, The Hunterian, University of Glasgow and the London Natural History Museum for access to fossil thin sections, and to the curators of these collections (S. Harris, K. Sherburn, N. Clark and P. Hayes) for their assistance. We thank T. Goral, C. Strullu-Derrien, C. Kirchhelle, I. Moore and I. Rahmen for assistance and advice for confocal imaging, segmentation and 3D reconstructions; J. Baker for photographic advice; and A. M. Hetherington, N. J. Hetherington and members of the Dolan laboratory for helpful comments and discussions.

Reviewer information *Nature* thanks P. Crane and P. Kenrick for their contribution to the peer review of this work.

Author contributions A.J.H. and L.D. designed the project. A.J.H. carried out the analyses. A.J.H. and L.D. wrote the paper.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0445-z.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0445-z.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to L.D.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Identifying apices of rooting axes of *A. mackiei*. To identify meristems at the apices of rooting axes of *A. mackiei*, we visually inspected 641 thin sections prepared from the Rhynie chert from the following collections: 11 from the collections of the School of Biology, University of St Andrews; 7 from the Oxford University Herbaria; 33 from the collections of the Manchester Museum, The University of Manchester; 299 from the Natural History Museum, London; 291 from The Hunterian, University of Glasgow.

Imaging of rooting apices. Photographs of thin sections were taken using a Nikon D80 with a 60-mm macro lens, set up on a copystand. Thin sections were lit from below with a lightbox and lit from above with aerial lights (Extended Data Figs. 1, 2b). High-resolution images of the apices were taken using Nikon Eclipse LV100ND (Figs. 1a–c, e, 2c, Extended Data Figs. 2a, c, 3a, c, 4a, b) and Olympus BX50 (Figs. 1d, 2a, Extended Data Figs. 3b, d, 4c) compound microscopes and a Leica M165 FC (Extended Data Fig. 4d) stereo microscope. To create high-definition images, multiple overlapping photographs were taken and combined to make a single image using AutoStitch[31].

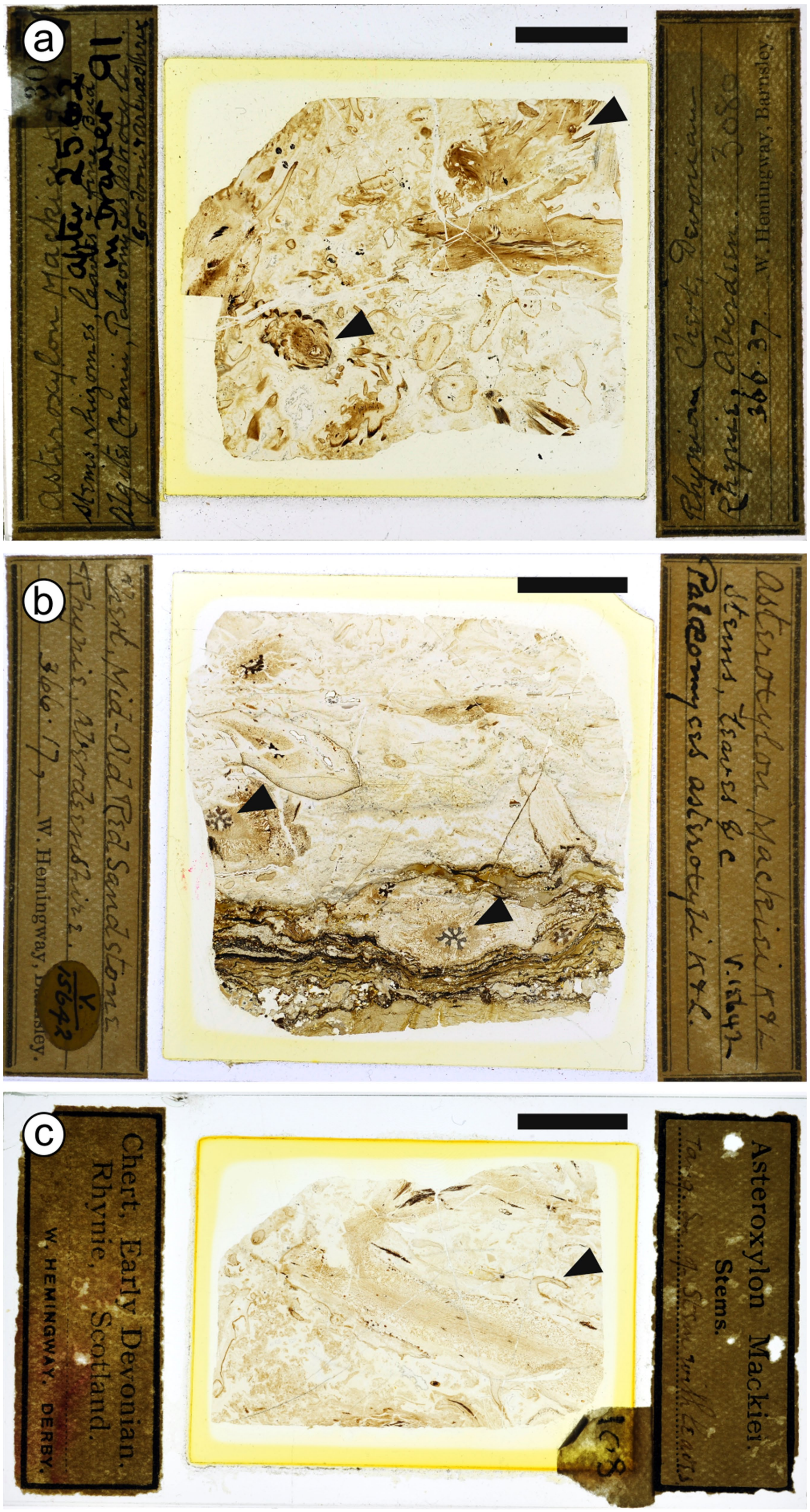
Confocal laser scanning microscopy was used to image meristems of rooting axes of *A. mackiei*. Confocal images were acquired with a Nikon A1-Si laser-scanning confocal microscope (Natural History Museum, London) (Figs. 1f, g, 2d) and a Leica SP5 confocal microscope (Department of Plant Sciences, University of Oxford) (Fig. 2b). The Nikon A1-Si laser-scanning confocal microscope was used with a 40× oil-immersion objective with a 1.3 numerical aperture and 29.37-μm pinhole. Autofluorescence of the sample was excited with a 561-nm and 640-nm laser and emission was collected with windows of 570–620 nm and 675–725 nm for each laser, respectively. The Leica SP5 confocal microscope was used with a 20× oil-immersion objective with a 0.7 numerical aperture and a 60.7-μm pinhole. Autofluorescence of the sample was excited with a 633-nm laser and emission was collected with a window of 645–800 nm.

Segmentation and visualization of *A. mackiei* meristem. Images were processed using FIJI[32]. Segmentation of the epidermal surface of the meristem was carried out in MorphoGraphX[33] (Supplementary Video 1). The three-dimensional model was visualized in Blender[34] (Fig. 3, Supplementary Video 2).

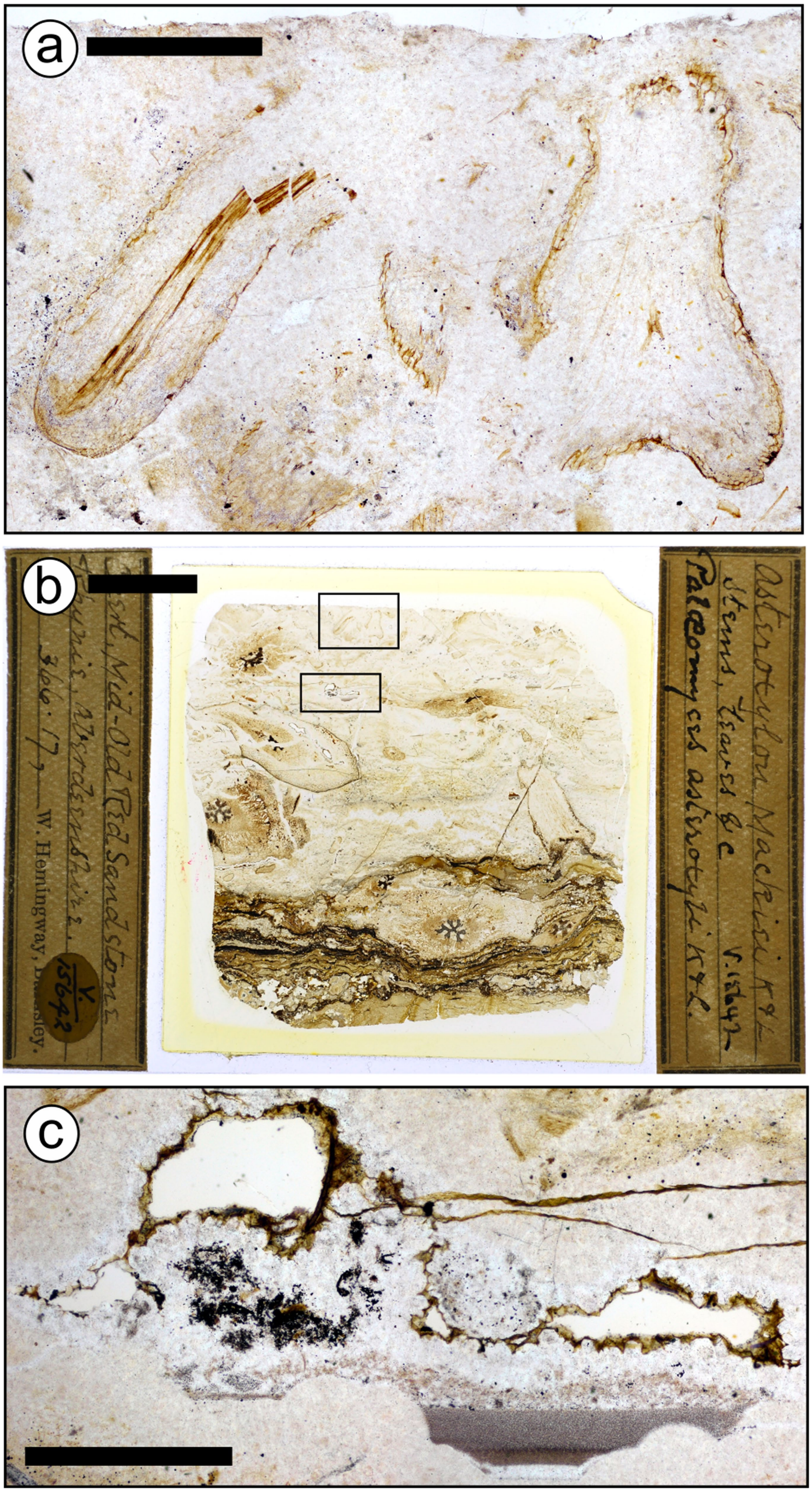
Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability. The three thin sections analysed in this study are housed in publically available collections: GLAHM Kid 3080, OXF 108 and NHMUK V.15642. The confocal laser scanning microscopy datasets generated are available from the corresponding author upon reasonable request. All other data supporting the findings of this study are included in the paper and its Extended Data and Supplementary Information.

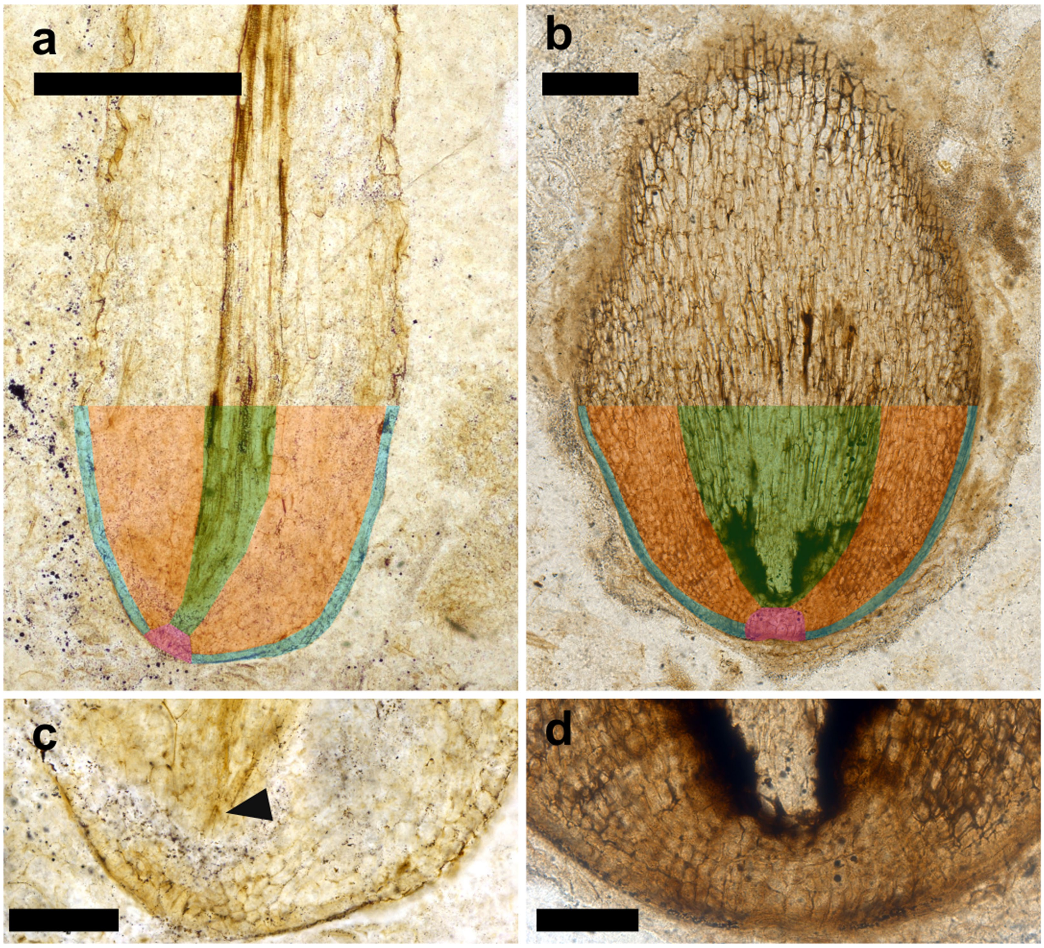
31. Brown, M. & Lowe, D. G. Automatic panoramic image stitching using invariant features. *Int. J. Comput. Vis*. **74**, 59–73 (2007).
32. Schindelin, J. et al. Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).
33. Barbier de Reuille, P. et al. MorphoGraphX: A platform for quantifying morphogenesis in 4D. *eLife* **4**, 05864 (2015).
34. Garwood, R. & Dunlop, J. The walking dead: Blender as a tool for paleontologists with a case study on extinct arachnids. *J. Paleontol*. **88**, 735–746 (2014).


a
W. Hemingway, Barnsley.
b
V.15642
W. Hemingway, Barnsley.
c
Chert, Early Devonian.
Rhynie, Scotland.
W. HEMINGWAY. DERBY.
Asteroxylon Mackiei.
Stems.
108

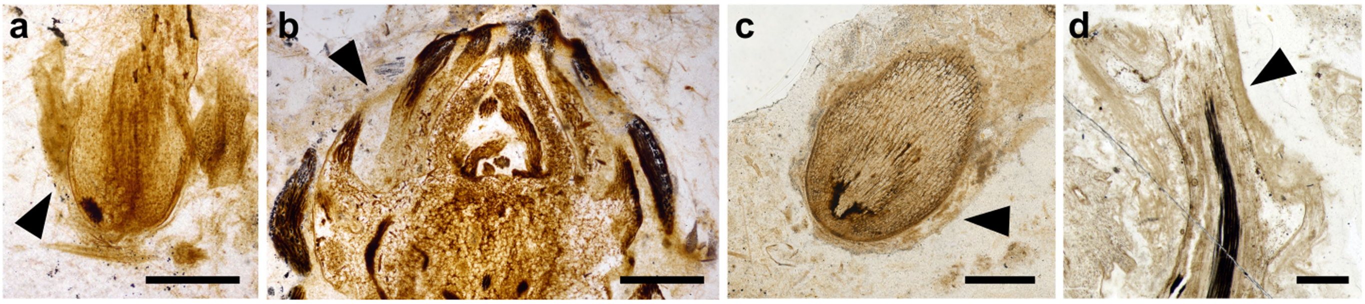
Extended Data Fig. 1 | Five root apices were found on three thin sections in which *A. mackiei* was the only plant species present. **a**–**c**, Diagnostic features of *A. mackiei* include the apices of leafy shoots (black arrowhead, **a**), star-shaped xylem (black arrowhead, **b**) and leaves (black arrowhead, **c**). **a**, GLAHM Kid 3080. **b**, NHMUK V.15642. **c**, OXF 108. Scale bars, 1 cm.


a
b
Mid-Old Red Sandstone
Rhynie, Aberdeenshire.
3.66.17,
W. Hemingway, Barnsley.
V.
15642
Asteroxylon Mackiei K&L
Stems, leaves &c V.15642
Palaeomyces asteroxyli K&L.
c

Extended Data Fig. 2 | *A. mackiei* rooting axes grew in the direction of the gravity vector. **a**–**c**, Positively gravitropic growth of two apices (**a**) was inferred from their orientation relative to sediment layers in both the growth substrate (dark brown and black bands at base of **b**) and a geopetally infilled void (**c**) preserved in the thin section. The position of both the apices and geopetally infilled void are highlighted with black boxes within the thin section (**b**). **a**–**c**, NHMUK V.15642. Scale bars, 1 mm (**a**, **c**), 1 cm (**b**).



Extended Data Fig. 3 | Fundamental tissues present in an apex of a rooting axis preserved after growth had finished, and a meristem of a rooting axis preserved during active growth. a, **b**, Root apices with fundamental tissue types colour-coded. Blue, epidermis; pink, promeristem; orange, cortex; and green, procambium. **c**, **d**, Magnified images of the apical regions of **a** (**c**) and **b** (**d**). The presence of differentiated vascular tissue (arrowhead, **c**) close to the tip of the apex indicates that this apex was not active at the time of preservation. By contrast, in **d** there is no differentiated vascular tissue. Instead, the apex is characterized by large numbers of cells and cell size gradually increases with distance from the tip, which indicates that the apex was active when fossilized. **a**, **c** NHMUK V.15642 (same specimen as illustrated in Fig. 1a), **b**, **c**, OXF 108 (same specimen as illustrated in Figs. 1d, 2a, b). Scale bars, 500 μm (**a**), 250 μm (**b**), 150 μm (**c**), 100 μm (**d**).



Extended Data Fig. 4 | Mulm coats the rooting axes and leafy shoots of *A. mackiei*. **a**–**d**, A thin layer of degraded organic material called mulm[27] (highlighted with arrowheads, **a**–**d**) coats both the rooting axes (**a**, **c**) and leafy shoots of *A. mackiei*. **a**, **b**, GLAHM Kid 3080. **c**, **d**, OXF 108. Scale bars, 500 μm (**a**, **c**, **d**), 1 mm (**b**).

LETTER

https://doi.org/10.1038/s41586-018-0460-0

Sensing with tools extends somatosensory processing beyond the body

Luke E. Miller[1,2,3]*, Luca Montroni[1,2], Eric Koun[1,2], Romeo Salemme[1,2,3], Vincent Hayward[4,5] & Alessandro Farnè[1,2,3,6]*

The ability to extend sensory information processing beyond the nervous system[1] has been observed throughout the animal kingdom; for example, when rodents palpate objects using whiskers[2] and spiders localize prey using webs[3]. We investigated whether the ability to sense objects with tools[4–9] represents an analogous information processing scheme in humans. Here we provide evidence from behavioural psychophysics, structural mechanics and neuronal modelling, which shows that tools are treated by the nervous system as sensory extensions of the body rather than as simple distal links between the hand and the environment[10,11]. We first demonstrate that tool users can accurately sense where an object contacts a wooden rod, just as is possible on the skin. We next demonstrate that the impact location is encoded by the modal response of the tool upon impact, reflecting a pre-neuronal stage of mechanical information processing akin to sensing with whiskers[2] and webs[3]. Lastly, we use a computational model of tactile afferents[12] to demonstrate that impact location can be rapidly re-encoded into a temporally precise spiking code. This code predicts the behaviour of human participants, providing evidence that the information encoded in motifs shapes localization. Thus, we show that this sensory capability emerges from the functional coupling between the material, biomechanical and neural levels of information processing[13,14].

Historically, researchers across scientific disciplines have focused almost exclusively on motor aspects of tool use[15,16], despite the fact that tools convey behaviourally relevant sensory information to the user when contacting a surface[17]. Indeed, tactile signals are critical for both hand and tool use as they provide information about objects that is unavailable in other modalities. One familiar example, first discussed in the seventeenth century by René Descartes[18], is the ability of people with visual impairments to perceive their environment through the tip of a cane[19] (Fig. 1a). Despite the passage of almost four centuries since the publication of Descartes' treatise, scientists have only recently begun investigating how handheld tools are used to sense the properties of the environment, such as the texture[6], position[8] and distance[4] of objects. Tool-mediated sensing is therefore a poorly understood aspect of the daily human experience.

It is unclear whether, when using a rod to manipulate an object, humans perceive where the object contacts its surface. Several distinct patterns may characterize perception during tool-mediated sensing (Fig. 1b). First, on the basis of a longstanding hypothesis (sensory distalization)[10], localization may be confined to the tip of the rod regardless of the actual contact location. Alternatively, localization may be projected to the proximal and distal regions of the rod (sensory projection), after the known deployment of spatial attention along a tool[11]. By contrast, we claim that handheld tools function as sensory extensions of the user's body (sensory embodiment). An embodied view[14] makes three clear predictions: (i) Humans can accurately localize contact along the entire surface of a tool; (ii) contact-related somatosensory information is represented in a coordinate system that is intrinsic to the surface of the tool (that is, it is tool-centred) and (iii) accurate sensing with a rod depends on the ability of the nervous system to anticipate and decode its structural dynamics. We combine multiple lines of evidence to characterize how humans use a tool to extend somatosensory processing.

When measuring tactile localization, it is common to require participants to indicate on a graphical representation of a limb where they were touched[20]. We adapted this task to the case of handheld wooden rods (see section 'Apparatus and impact localisation task' of the Methods; Fig. 1b, Extended Fig. 1a). Participants localized impacts (seven distinct locations) on a downsized graphical representation of a rod, a task that requires mapping tactile signals within a coordinate system that is intrinsic to the space of the tool and not the external space it occupies. Localization occurred after contact with an object through either self-generated action (active sensing) or passive reception of impact (passive sensing). Comparing these sensing modes allowed us to infer the roles of sensory and motor signals in the perception of impact location.

In all experiments, we used affine regression to assess localization performance. The slope reflects the perceived separation between landmarks and is therefore a measure of performance. Our analysis compared estimated slopes to random ($\text{slope} = 0.25$) and ideal localization ($\text{slope} = 1$). In doing so, we could test each hypothesis of tool-mediated sensing as only sensory embodiment predicts accurate localization. We further tested the models of sensory embodiment and projection (Fig. 1b) by comparing the Akaike information criterion of each model (ΔAIC) for each participant.

In an initial experiment ($n = 10$) participants used both sensing modes (order counterbalanced) after a five-minute familiarization phase. Active sensing produces a rich array of motor (for example, efference copies) and transient sensory signals (for example, cutaneous vibrations) that could be used to extract impact location. Localization during active sensing was highly accurate ($\text{slope} = 0.93 \pm 0.09$; one-sample t-test versus random: $P < 0.001$; Fig. 1d), reflecting near-ideal performance (versus ideal, $P = 0.46$). Compared to active sensing, passive sensing largely removes motor-related signals while preserving most of the sensory signals. Participants consistently underestimated the distances between impact locations ($\text{slope} = 0.57 \pm 0.04$), as expected when informative cues are removed[21]. Nevertheless, performance during passive sensing was still accurate (versus random, $P < 0.005$), although it was substantially lower than in the active condition (paired t-test, $P < 0.005$). These results clearly favour sensory embodiment over distalization.

Follow-up experiments provided a more complete picture of extended sensing (Extended Table 1). To demonstrate the robustness of our initial findings, we replicated them for both active (Extended Fig. 2a) and passive sensing (Extended Fig. 2b). In a fourth experiment, we found that participants localized contact on a rotated drawing with equal accuracy as when the drawing was displayed parallel to the rod (paired t-test, $P = 0.44$; correlation, $r = 0.89$; Extended Fig. 2c–e), providing further evidence that users internally represent the rod in tool-centred coordinates.

---

[1] Integrative Multisensory Perception Action & Cognition Team—ImpAct, Lyon Neuroscience Research Center, INSERM U1028, CNRS U5292, Lyon, France. [2] University of Lyon 1, Lyon, France. [3] Hospices Civils de Lyon, Neuro-immersion, Lyon, France. [4] Sorbonne Université, Institut des Systèmes Intelligents et de Robotique (ISIR), Paris, France. [5] Centre for the Study of the Senses, School of Advanced Study, University of London, London, UK. [6] Center for Mind/Brain Sciences, University of Trento, Rovereto, Italy. *e-mail: luke.miller@inserm.fr; alessandro.farne@inserm.fr

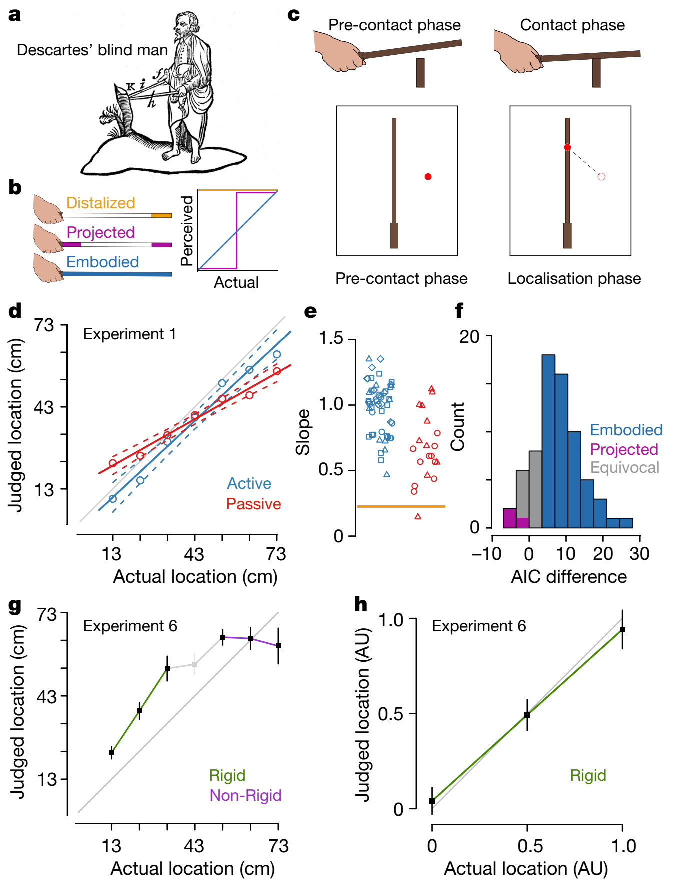


Fig. 1 | Methods and behavioural results. a, Descartes' blind man using two canes to triangulate the distance of an object. Image has been adapted from an illustration made by Descartes with permission from Getty Images. **b**, Hypothetical models of localization. Left, coloured regions correspond to where contact is felt, as predicted by each model; white regions are perceptually invisible to the user. Right, expected patterns of results when modelling perception as a function of actual impact location. Unlike the other two hypotheses embodiment predicts that impact localization will manifest as a linear function (blue line) for the entire body of the tool, mirroring what we observe for tactile localization on a body part. **c**, Three phases composed each trial, including object contact (top) and localization with the cursor (red circle; bottom row). **d**, Group-level affine regressions for experiment 1 ($n = 10$). Dashed lines correspond to the 95% confidence interval. The grey line corresponds to the equality line. These results were replicated in experiments 2–5. **e**, Slope for each participant from experiments 1–5 ($n = 60$). The prediction of the distalization model (that is, chance performance) is shown by the orange line. **f**, Model comparisons (AIC difference: projected − embodiment) for every participant in experiments 1–5 ($n = 60$). The majority of comparisons favoured sensory embodiment (blue) with substantially less either favouring sensory projection (purple) or providing equivocal evidence (grey). **g**, Experiment 6 ($n = 10$), participants failed to differentiate between distinct locations when contact was made with the non-rigid portion of the rod (purple; unpredictable dynamics). By contrast, localization had a positive slope when contact was made with the rigid portion (green; predictable dynamics). **h**, Judgments shown in **g** were above the equality line because participants overestimated the contribution of the rigid portion to the overall length of the rod (76.3% ± 4.8). Judgments overlapped with the equality line (grey) when normalized to the felt rigid-to-length ratio. AU, arbitrary units. Error bars in **g** and **h** correspond to the s.e.m.

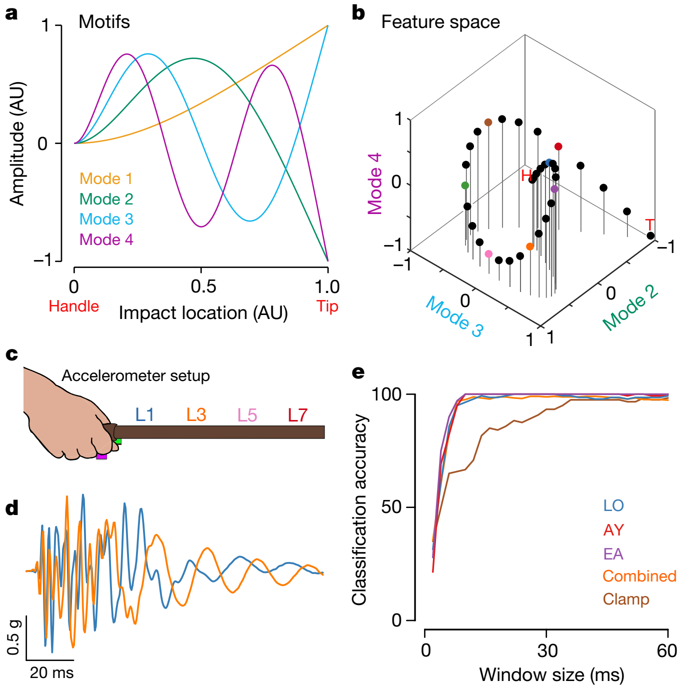


Fig. 2 | Vibratory motifs emerge rapidly during extended location sensing. a, The amplitude of the first through fourth resonant modes as a function of impact location (see Supplementary Data). Impact at each location along the rod produces a unique combination of modal amplitudes, which we term vibratory motifs. **b**, Hypothetical feature space of motifs constructed from the modes shown in **a**. The geometry of this space is, theoretically, identical for all uniform rods used by humans. Impact at any location (circles) from the handle (H) to the tip (T) produces a motif that is situated in a unique position within the feature space. The coloured dots correspond to the impact locations used in our experiment. **c**, Vibrations on the tool (green) and index finger (purple) were measured simultaneously after impacts at seven distinct locations (L1–L7). **d**, Motifs from the dataset of participant EA for L1 and L3. Similar patterns were observed for participants LO and AY. **e**, The accumulation of location information within motifs was rapid for each participant and when all datasets were combined (orange), demonstrating that the motifs of each participant share location-specifying features. All cases rivalled what was observed for vibrations when the tool was fully clamped (brown; see pilot experiment in Supplementary Data section 3), suggesting that the information encoded by motifs is strengthened by the biomechanical properties of the hand.

Finally, we probed the importance of sensorimotor predictions for extended sensing[9]. In a fifth experiment, we found that accurate location sensing did not depend on prior experience wielding the rod (Extended Fig. 2f), probably because the adult nervous system has internal mechanisms for predicting the structural dynamics of wooden rods. To link these dynamics to impact location, however, the nervous system must assume that the rod is uniformly rigid (see Supplementary Data). We violated this assumption in a sixth experiment by requiring participants to wield a rod whose handle and first half was rigid (wood) and second half was non-rigid (foam; Extended Fig. 1b). Prior to the localization task, users held the rod by its handle but never saw or wielded it. The structural dynamics of its wooden half were therefore predictable, whereas they were unpredictable for the foam half (Extended Fig. 2g). Accurate localization was impossible when contact was on the foam half (Fig. 1g). By contrast, users accurately rescaled their localization on the wooden half (Fig. 1g–h), demonstrating that the nervous system adapts its expectations about dynamics when novel tools violate the rigidity assumption.

Our behavioural experiments converged to a similar conclusion. The performance of almost every participant was above chance (Fig. 1e), ruling out the sensory distalization model. Furthermore, sensory embodiment was a significantly better model than sensory projection in the vast majority of datasets (54/70 versus 3/70; mean ΔAIC = 7.62 ± 0.76; Fig. 1f). These results are in line with the predictions of sensory embodiment and provide strong evidence that tools function as sensory extensions of the body.

The high accuracy observed in each experiment demonstrates that sensory embodiment is largely independent of sensing mode.

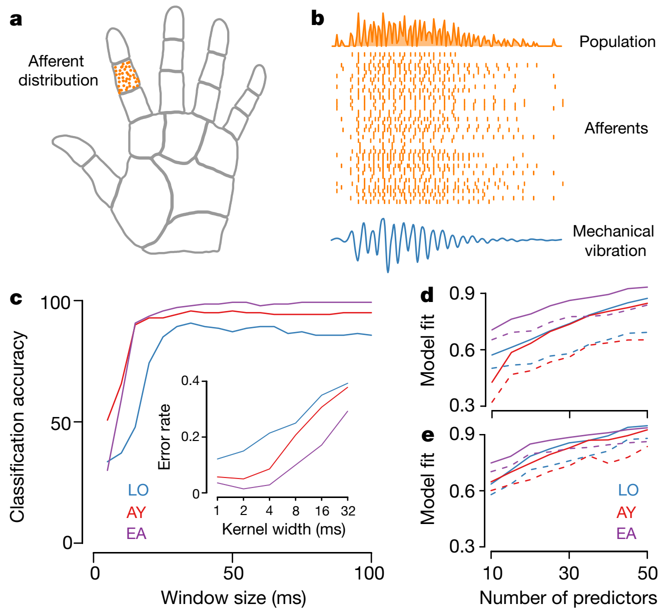


Fig. 3 | Impact location is encoded in the spike timing of tactile mechanoreceptors in the hand. a, The distribution of 42 simulated Pacinian mechanoreceptors in the middle phalanx of the index finger. This image has been produced using TouchSim[12]. **b**, Mechanical vibration (first 150 ms after impact) on a randomly chosen trial from the dataset of participant LO. Vibrations led to temporally precise response patterns across the individual afferents (orange raster plot) and the population response (top). **c**, We could accurately decode impact location from the population spike-timing. (inset) The classification error rate increased notably for kernel widths greater than four milliseconds. **d–e**, Model fits of spikes (**d**) and motifs (**e**) for the behaviour of each participant as a function of the number of predictors (140 data points per test). Solid lines correspond to R^2 and dashed lines correspond to predictive R^2.

The differences between active and passive sensing do, however, have implications for the information used during extended sensing. Superior performance during active sensing (Extended Fig. 2g) suggests that human tool users utilize information encoded in both sensory and motor signals[8]. Nevertheless, the notably high accuracy during passive sensing suggests that sensory signals, alone, encode a substantial portion of spatial information.

We next asked how a tool communicates impact location to the hand. Tactile mechanoreceptors in the human hand[22] are highly sensitive to the cutaneous vibrations[23] elicited during object manipulation. According to the Euler–Bernoulli beam theory, a rod resonates according to well-defined modes when contacting an object (Supplementary Data section 1). Importantly, the relative amplitude and phase of each mode depends almost exclusively on where contact occurs along the rod, relative to its length (Fig. 2a, Extended Fig. 3). The modal response thus encodes an invariant signal of location[24], suggesting that rods are a highly robust means to extend somatosensory processing provided that users can predict key aspects of their material and geometry (experiments 5–6). We therefore hypothesized that during tool-extended sensing, a rod mechanically transduces impact location into vibratory motifs (Fig. 2a–b) that are decoded by the somatosensory system.

We recorded vibrations on the handle of a rod and the index finger of three human participants while they performed either passive (participants LO and AY) or active (participant EA) extended sensing (Fig. 2c, Extended Fig. 1c). Behavioural results for each participant were within the range observed previously (Extended Fig. 2i). Impact at each location produced distinct vibratory motifs (Fig. 2d) that were highly consistent across trials (Extended Fig. 4). Support vector classification found that the location-specific pattern of each motif emerged extremely rapidly, with classification accuracy reaching 90% within approximately 8 ms after impact (Fig. 2e). For each motif, high classification accuracy was also found for the patterns of its phase-locked temporal encoding (Extended Fig. 5a–c), a subspace of the modal response for which analysis is computationally efficient and neurophysiologically plausible (Supplementary Data section 2). Similar results were found for cutaneous vibrations (Extended Fig. 6). Vibratory motifs are therefore an informationally rich signal that is likely exploited by the user during tool-extended sensing.

Our findings thus far provide evidence that the modal response of a rod reflects the initial transduction of information about impact location. This finding is comparable to object localization during whisking by rats, in which it has been suggested that information is initially processed pre-neuronally by the mechanics of the whiskers[2]. If vibratory motifs are to guide behaviour, mechanoreceptors in the hand must transduce them into neural response patterns that preserve the location-specifying information, a large portion of which is carried by the phase-locked encoding of a motif (Extended Fig. 5). Pacinian mechanoreceptors are good candidates for this transformation given their broad frequency tuning[25], temporally precise spiking[26] and phase-locked response properties[27]. They have been proposed to have an important role in encoding vibrations transmitted through handheld objects[5]. We leveraged a biologically plausible skin–neuron model of the hand (TouchSim)[12] to simulate how impact location may be re-encoded by mechanoreceptors during extended sensing.

We simulated the spiking responses of a population of 42 Pacinian mechanoreceptors (Fig. 3a) to the mechanical vibrations described above. Simulated spikes showed a temporally precise phase-locked relationship with the motifs (Fig. 3b, Extended Fig. 5a–b). The population response started 6.2 ± 0.8 ms after impact, suggesting that pre-neuronal and neuronal representations of object location co-evolve in time. Importantly, substantial location information emerged in the response of the population within approximately 25 ms (Fig. 3c) and was dependent upon the millisecond precision of spiking (Fig. 3c, inset; Extended Fig. 7). This time course is in line with the responses of mechanoreceptors during object manipulation with the hand[28], suggesting that the nervous system can extract sensory information from a tool with a similar speed as the body itself. Pacinian mechanoreceptors may therefore reduce the dimensionality of motifs while preserving information about impact location (see Supplementary Data section 2).

The trial-by-trial judgements of each participant were predicted well from the corresponding simulated population spiking pattern (multivariate regression with 40 predictors: all $R^2 > 0.8$; all predictive $R^2 > 0.6$; Fig. 3d, Extended Table 2), which provided significantly better fits than firing rate (Extended Fig. 8). Similar results were found when motifs were used as predictor variables instead (Fig. 3e, Extended Fig. 5d). Furthermore, both spike timing and motifs accurately predicted the trial-by-trial directional errors (all $R^2 > 0.75$; all predictive $R^2 > 0.55$), further underscoring the importance of these factors for behaviour.

In conclusion, we have provided behavioural, mechanical and computational evidence that humans utilize tools as sensory extensions of the body. Our results suggest that the human nervous system contains finely tuned mechanisms for decoding vibratory motifs, including sensorimotor internal models that can anticipate the structural dynamics of rods[9,29]. Object localization with a rod therefore represents an important human model of extended sensing[1], in which somatosensory processing is distributed across the mechanical response of the tool, the biomechanics of the extremities, and the neural circuits of the sensorimotor system[13]. Morphological and neural changes in the hand during human evolution[30] may reflect selection pressure to maximize the functional coupling and embodiment of tools[16,17].

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0460-0.

Received: 25 September 2017; Accepted: 2 July 2018;
Published online 12 September 2018.

1. Clark, A. & Chalmers, D. The extended mind. *Analysis* **58**, 7–19 (1998).
2. Bagdasarian, K. et al. Pre-neuronal morphological processing of object location by individual whiskers. *Nat. Neurosci*. **16**, 622–631 (2013).
3. Japyassú, H. F. & Laland, K. N. Extended spider cognition. *Anim. Cogn*. **20**, 375–395 (2017).
4. Chan, T. C. & Turvey, M. T. Perceiving the vertical distances of surfaces by means of a hand-held probe. *J. Exp. Psychol. Hum. Percept. Perform*. **17**, 347–358 (1991).
5. Brisben, A. J., Hsiao, S. S. & Johnson, K. O. Detection of vibration transmitted through an object grasped in the hand. *J. Neurophysiol*. **81**, 1548–1558 (1999).
6. Klatzky, R. L. & Lederman, S. J. Tactile roughness perception with a rigid link interposed between skin and surface. *Percept. Psychophys*. **61**, 591–607 (1999).
7. Yamamoto, S. & Kitazawa, S. Sensation at the tips of invisible tools. *Nat. Neurosci*. **4**, 979–980 (2001).
8. Saig, A., Gordon, G., Assa, E., Arieli, A. & Ahissar, E. Motor-sensory confluence in tactile perception. *J. Neurosci*. **32**, 14022–14032 (2012).
9. Kilteni, K. & Ehrsson, H. H. Sensorimotor predictions and tool use: Hand-held tools attenuate self-touch. *Cognition* **165**, 1–9 (2017).
10. Katz, D. *The World of Touch* (Erlbaum, Hillsdale, 1989).
11. Holmes, N. P., Calvert, G. A. & Spence, C. Extending or projecting peripersonal space with tools? Multisensory interactions highlight only the distal and proximal ends of tools. *Neurosci. Lett*. **372**, 62–67 (2004).
12. Saal, H. P., Delhaye, B. P., Rayhaun, B. C. & Bensmaia, S. J. Simulating tactile signals from the whole hand with millisecond precision. *Proc. Natl Acad. Sci. USA* **114**, E5693–E5702 (2017).
13. Chiel, H. J. & Beer, R. D. The brain has a body: adaptive behavior emerges from interactions of nervous system, body and environment. *Trends Neurosci*. **20**, 553–557 (1997).
14. Makin, T. R., De Vignemont, F. & Faisal, A. A. Neurocognitive barriers to the embodiment of technology. *Nat. Biomed. Eng*. **1**, 0014 (2017).
15. Johnson-Frey, S. H. The neural bases of complex tool use in humans. *Trends Cogn. Sci*. **8**, 71–78 (2004).
16. Martel, M., Cardinali, L., Roy, A. C. & Farnè, A. Tool-use: an open window into body representation and its plasticity. *Cogn. Neuropsychol*. **33**, 82–101 (2016).
17. Miller, L. E., Longo, M. R. & Saygin, A. P. Tool morphology constrains the effects of tool use on body representations. *J. Exp. Psychol. Hum. Percept. Perform*. **40**, 2143–2153 (2014).
18. Descartes, R. in *Œuvres et lettres* (Gallimard, Paris, 1937).
19. Schenkman, B. N. & Jansson, G. The detection and localization of objects by the blind with the aid of long-cane tapping sounds. *Hum. Factors* **28**, 607–618 (1986).
20. Mancini, F., Longo, M. R., Iannetti, G. D. & Haggard, P. A supramodal representation of the body surface. *Neuropsychologia* **49**, 1194–1201 (2011).
21. Kersten, D., Mamassian, P. & Yuille, A. Object perception as Bayesian inference. *Annu. Rev. Psychol*. **55**, 271–304 (2004).
22. Johnson, K. O. The roles and functions of cutaneous mechanoreceptors. *Curr. Opin. Neurobiol*. **11**, 455–461 (2001).
23. Shao, Y., Hayward, V. & Visell, Y. Spatial patterns of cutaneous vibration during whole-hand haptic interactions. *Proc. Natl Acad. Sci. USA* **113**, 4188–4193 (2016).
24. Hayward, V. in *Human Haptic Perception: Basics and Applications* (ed. Grunwald, M.) 381–392 (Birkhäuser, Basel, 2008).
25. Bell, J., Bolanowski, S. & Holmes, M. H. The structure and function of Pacinian corpuscles: a review. *Prog. Neurobiol*. **42**, 79–128 (1994).
26. Mackevicius, E. L., Best, M. D., Saal, H. P. & Bensmaia, S. J. Millisecond precision spike timing shapes tactile perception. *J. Neurosci*. **32**, 15309–15317 (2012).
27. Talbot, W. H. & Mountcastle, B. The sense of flutter–vibration: comparison of the human capacity with response patterns of mechanoreceptive afferents from the monkey hand. *J. Neurophysiol*. **31**, 301–334 (1968).
28. Johansson, R. S. & Birznieks, I. First spikes in ensembles of human tactile afferents code complex spatial fingertip events. *Nat. Neurosci*. **7**, 170–177 (2004).
29. Imamizu, H. et al. Human cerebellar activity reflecting an acquired internal model of a new tool. *Nature* **403**, 192–195 (2000).
30. Young, R. W. Evolution of the human hand: the role of throwing and clubbing. *J. Anat*. **202**, 165–174 (2003).

Acknowledgements We thank F. Volland for his help constructing the experimental setup; B. Miller and A. Schork for their statistical advice; and L. Quadt, A. Roy, E. Leonardis and K. Kilteni for their feedback on an earlier version of the manuscript. This work was supported by an FRM postdoctoral fellowship to L.E.M., ANR-16-CE28-0015 Developmental Tool Mastery to A.F. and V.H., a Leverhulme Trust Visiting Professorship Grant to V.H. and IHU CeSaMe ANR-10-IBHU-0003, Defi Auton Sublima and the James S. McDonnell Scholar Award to A.F. All work was performed within the framework of the LABEX CORTEX (ANR-11-LABX-0042) of Université de Lyon.

Reviewer information *Nature* thanks S. Bensmaia and G. Stanley for their contribution to the peer review of this work.

Author contributions L.E.M., V.H. and A.F. conceived the study and designed the experiments. L.E.M. and L.M. collected and analysed the behavioural experiments. E.K. and R.S. constructed equipment for the vibration experiment and helped process the data. L.E.M., V.H. and A.F. designed and analysed the neuronal modelling and vibration experiments. V.H. developed the theoretical framework presented in the Supplementary Data. L.E.M., V.H. and A.F. wrote the paper. All authors approved the final version of the paper.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0460-0.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0460-0.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to L.E.M. or A.F.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Participants. Sixty-three participants in total completed our behavioural experiments: ten in experiment 1 (nine right-handed, four males, 27.2 ± 3.1 years of age), ten in experiment 2 (ten right-handed, five males, 27.8 ± 2.6 years of age), ten in experiment 3 (nine right-handed, four males, 26.5 ± 2.5 years of age), twenty in experiment 4 (all right-handed, twelve males, 22.6 ± 2.1 years of age), and ten in experiment 5 (nine right-handed, seven males, 29.1 ± 2.7 years of age); the same ten participants from experiment 5 also completed experiment 6 (order counterbalanced). A further three right-handed participants completed a more in-depth psychophysical experiment: LO (female; 24 years of age), AY (female; 25 years of age) and EA (male; 20 years of age). All participants had normal or corrected-to-normal vision and no history of neurological impairment. Every participant gave informed consent to participate in the experiment. The study was approved by the ethics committee (CPP SUD EST IV, Lyon, France).

Apparatus and impact localization task. The setup for each behavioural experiment (Extended Fig. 1a) was as follows: participants were either seated comfortably in a cushioned chair (experiments 1–4 and 7) or standing in a comfortable posture (experiments 5 and 6). Their right arm was placed in a padded arm rest and hidden from view with a long occluding board. An LCD screen (47 × 30 cm) lay backside down nine cm from the edge of the table and in the lengthwise orientation (centred on the participant's midline). This orientation allowed us to display a long computer drawing of the tool that was viewed by the participants during the task (see below).

Several types of wooden rods were used in the experiments. Their details are as follows: In experiments 1–3 and 7, the wooden rod (Byron & Byron; model: tiara draw rod) had a 10-cm handle and an 83-cm body (0.6 cm radius). In experiment 4 the wooden rod had a 12-cm handle and a 60-cm body (0.75 cm radius). In experiment 5 the wooden rod had a 12-cm handle and an 83-cm body (0.75 cm radius) that was insulated from handle-to-tip with a lightweight foam covering. In experiment 6, we used a hybrid tool that was approximately half rigid and half non-rigid. The rigid portion of the rod was wood with 12-cm handle and a 38-cm body (0.75 cm radius). As before, the rod was insulated with a lightweight foam material that covered both the wooden portion and extended a further 45-cm, thus forming a non-rigid portion. The dynamics of each half were therefore very different, with only the rigid portion being predictable to the participant, given that it was the only portion they were exposed to before wielding (see below). The rod was held in the participant's right (experiments 1–4 and 7) or dominant hand (experiments 5–6) and was always fully blocked from view by the occluding board.

The task in each experiment was to localize where an object contacted the surface of the tool by pointing to the corresponding location on the graphical representation of the tool. This drawing was scaled to 40% of the actual dimensions of the rod; it began 14.5 cm from the edge of the table and was raised 4 cm above the table surface. A red cursor (0.2 cm radius) was placed 10 cm to its right. The drawing was displayed in parallel with the actual rod in all experiments except experiment 4. In this experiment, the drawing of the rod was rotated 90° counter-clockwise on half of the trials (randomly interleaved with the parallel drawings). This manipulation allowed us to more rigorously characterize how the rod is internally represented. If localization performance is independent of drawing orientation, it would provide strong evidence that the rod is represented in a tool-centred coordinate system. Indeed, debriefing interviews after the experiment found that most participants found both orientations equally challenging. Although it cannot be completely ruled out that localizing on the rotated drawing was done using explicit mental rotation, it is unlikely as the majority of participants reported that the task required little effort or attention to switch between response modes (that is, upright and rotated).

In most experiments (1–3 and 5–7) there were seven distinct impact locations (that is, landmarks), ranging from 13 to 73 cm from the handle (by steps of 10 cm). In experiment 4 there were six distinct landmarks, ranging from 10 to 60 cm from the handle (by steps of 10 cm). This information was never given to participants. There were 10 trials per landmark in the first six experiments (70 trials total) and 20 trials per landmark in the final experiment (140 trials total). The specific landmark for each trial was chosen pseudo-randomly.

The object used to contact the rod in the first six experiments was a wooden block (3 cm width) with a tightly wrapped wool padding (~3.5 cm contact area) to minimize the sound of impact. The objects used in the final experiment were a narrow plastic stick (0.2 cm width) for two participants (LO and AY) and a wooden block (3 cm width) for one participant (EA). To further minimize auditory cues during the task, pink noise was played over headphones. Pilot studies confirmed that residual auditory cues were minimal and uninformative about impact location.

Our experiments tested two distinct types of sensing impact location: active sensing (experiments 1, 2 and 4–7) and passive sensing (experiments 1, 3 and 7). During active sensing, the participant wielded the tool into contact with the object. They were instructed to briefly tap the object a single time and not rest the tool on it, meaning that active wielding required both self-generated active and reactive movements. During passive sensing, the experimenter manually moved the object into contact with the tool, transferring mechanical energy into the body of the tool that was passively received by the participant. The only tool movements in this sensing mode were passively reactive. In experiment 1, in which each participant sensed both actively and passively, each sensing mode occurred in distinct blocks (order counterbalanced across participants).

Experiments 1–4 and 7 began with a sensorimotor familiarization session. During this session, participants were told to explore how the tool felt to contact a surface (padded edge) at different locations on the tool. Emphasis was placed on the vibratory aspect of the contact. Participants had full auditory and visual feedback throughout this familiarization session, which was self-paced and lasted for 5 min. Next, these participants were given a brief practice session (seven trials) with the localization task to familiarize themselves with the trial structure (see below). The actual localization task commenced after this practice session.

By contrast, there were no sensorimotor familiarization or practice sessions in experiments 5 and 6. Instead, participants were handed the rod only when the experiment was ready to commence, beginning the actual localization task roughly two minutes after they were given the rod. At no time did they ever see or hear the rod. Only somatosensory feedback about the material and geometry of the rod was available. This allowed us to test whether specific familiarity with the rod was necessary for accurate localization, or whether participants could use general knowledge about the dynamics of rods to sense impact location (see Supplementary Data).

The trial structure in each experiment was as follows: in the 'pre-contact phase', participants sat with their left hand on a trackball, their trunk centred on the drawing of the tool, and their right hand holding the tool situated behind the occluding board. A red cursor (circle, 0.2 cm radius) was placed 10 cm to the right of the tool drawing. A go cue (tap on the right shoulder) indicated the start of the 'contact phase'. They therefore had to either move the tool into contact with the object (active sensing; experiments 1, 2 and 4–7) or wait passively to sense the impact (passive sensing; experiments 1, 3 and 7), which occurred approximately one second after the go cue. In the final phase of the trial—the 'localization phase'—participants made their judgment about impact location. This was done by moving the cursor onto the graphical representation of the tool corresponding to the impact location and clicking the mouse. The drawing then briefly disappeared for 500 ms before reappearing in the same position, indicating the beginning of the next trial. Participants never received feedback about their performance or the correct impact location.

Participants used a hybrid rod in experiment 6 (see above for details). At the end of the experiment, participants used a tape measure to report the felt length of the rigid and non-rigid portion separately. A debriefing interview after the experiment found that participants were confident in the length and material of the wood portion. Given that only somatosensory information about the material from the handle was available to participants in experiment 6, as far as they knew the rod was only made of wood—after the experiments, interviews confirmed that this was indeed the belief of every participant at the beginning of the experiment. They were therefore unsure of what the non-rigid portion of the rod was made of and used a purely cognitive strategy to infer that contact on its surface must be farther than the tip of the wood portion.

Vibration recording apparatus and experiments. We hypothesized that information about impact location would be encoded by the modal response of the rod when contacted. Vibrations were recorded using miniature tri-axis analogue accelerometers (Analogue Devices; model ADXL335). These devices have low mass (40.0 mg), wide frequency bandwidth (0–1,600 Hz in X and Y; 0–550 Hz in Z), and high dynamic range (−3.6 to 3.6 g). The signals were digitized with a 14-bit resolution and sampled at a frequency of 2.5 kHz over a five-second window using a data acquisition device (National Instruments; model USB-6009). Recording was restricted to the Z-axis as this contained the bulk of the impact response. Adhesive tape was used to keep the accelerometers in place on the surface of interest. For participants LO and AY, a single accelerometer was attached to the handle of the tool and one to the middle phalanx of the index finger (D2m; dorsal surface) and vibrations were recorded from each surface simultaneously. For participant EA, we decided to focus exclusively on mechanical vibrations. Therefore, a single accelerometer was attached to the base of the tool shaft. All data was recorded and processed using Matlab 2016b (The MathWorks).

Vibration experiment: tool held in the hand. The goal of this experiment was to investigate whether location information was encoded in mechanical vibrations during tool sensing (Extended Fig. 1c). Vibrations were recorded while participants passively (LO and AY) or actively (EA) sensed impact location. The behavioural task was identical to the behavioural experiments with the exception that there were twenty trials for each of the seven locations. In the passive sensing condition, each participant was trained to hold the tool as still as possible and with a similar grip force across trials. Great care was taken by the experimenter to ensure that this object contacted the tool with a similar force across trials. Recordings made directly from the tool occurred at its base and approximately two centimetres proximal from the beginning of the handle. The index finger was placed directly behind this accelerometer. In the active sensing condition, EA was trained to hold the tool with a stable grip force when resting and to contact the object with as similar of force as possible across all trials. Vibrations were recorded from a single accelerometer placed at the base of the tool shaft 1 cm distal to the handle. We set a minimum ten-second interval between each trial to ensure that all vibrations had fully dampened.

Signal processing. We used the following steps to process the vibrations recorded on each trial of every experiment. First, we converted the signal from voltage into a unit of acceleration, $g = 9.81$ m/s^2. Second, we filtered the signal using a zero-phase finite impulse response (FIR) filter with a bandpass between 100–600 Hz. This specific bandpass was chosen because it isolated the second through fourth modes (Fig. 2a–b), it served to remove most of the motor components from the accelerometer readings, and pilot studies indicated that it best captured location information. Third, we aligned the onset of vibrations to the zero time point and cut the signal with a window of −100–300 ms after impact. Fourth, we used mean subtraction to remove the baseline from the signal.

Skin–neuron model. We used a biologically plausible computational skin-neuron model[12] (TouchSim) to simulate putative mechanoreceptor responses in the hand during tool-mediated sensing. Matlab code to implement the model is freely available online (http://bensmaialab.org/download/). The reader should refer to the original article for an in-depth treatment of the methods. In brief, TouchSim simulates the spiking response and the spatial distribution of three different types of tactile afferents in the volar region of the hand: slow-adapting type one, rapidly adapting and Pacinian. The responses for each afferent type to a given stimulus are determined in two sequential stages: First, both the local deformation of the skin (static component) and the propagation of surface waves across the hand (dynamic component) are estimated. Second, this spatiotemporal pattern of stresses is used as input into afferent-specific integrate-and-fire models to determine the spiking response to the given stimulus. The model parameters of each afferent were derived from spiking data obtained in monkeys. The simulated responses of afferents closely match the known spiking responses (both millisecond precise spike-timing and rate) of actual afferents to various classes of stimuli (for example, vibrations, edges, textured surfaces).

We modelled the responses of a population of Pacinian afferents to the mechanical vibrations measured in the vibration experiment. The population was composed of 42 realistically distributed afferents in D2m; we confined our population to D2m as this was adjacent to where vibrations were measured from the handle. The response profile of each afferent was binned with a 1-ms resolution, allowing us to investigate its spike-timing with high temporal precision. To more realistically simulate afferents, the temporal profile of their responses included stochastic noise and considered known mechanical and spiking delays. Population-level spiking was created by summing the spiking of each individual afferent. The rate code for each afferent was calculated as its firing rate over time.

The mechanical vibration dataset for each participant (LO, AY and EA) was used as input for the model. Double integration with the trapezoidal method was used to convert acceleration (m/s^2) into displacement (μm). A zero-phase FIR filter with a high-pass at 80 Hz removed any accumulation of low-frequency error in the signal during this process. The time window of each vibration was restricted to 0–150 ms after impact. The modelled stimulus area had a radius of 7 mm and an indentation of 2 mm and was centred on D2m. This size roughly approximates the actual stimulus area when grasping a tool. Virtually identical results were found for a stimulus area with a radius of 3 mm (data not shown).

Data analysis. *Behavioural analysis.* To assess the ability of a participant to localize impact location on a tool, the mean localization judgement for each of the landmarks was fit with a least-squares affine regression. The decision to fit our data with a linear model was made before collecting data and because nonlinear models did not provide better fits in pilot experiments. The judged impact location on the drawing was converted from pixels into centimetres (that is, tool space) and was modelled as a function of actual impact location. The slope of the regression was used as our main measure of localization ability, although identical results are found when analysing the intercept.

To first assess whether the slope was greater than expected by chance (as is predicted by the sensory distalization model), we created a bootstrapped distribution of the slopes of regressions fit to simulated datasets of random guesses (100,000 simulations). The upper 95% confidence interval of this distribution (that is, 5,000th highest ranked element) was a slope of 0.25; a one-sample *t*-test was used to compare the actual slopes to this value. Next, to assess whether performance was 'near ideal', we used a one-sample *t*-test to compare the actual slopes to 1 (that is, the equality line) and the actual intercepts to 0. Paired *t*-tests were used in experiments 1 and 4 to compare performance in the within-subject conditions. All statistical tests were two-sided.

We further fit every participant's data (from the first five experiments) with the predicted pattern of results from the sensory projection model (Fig. 1b, right inset). For experiment 4, this was modelled as a proximal judgment of 10 cm for contact at the first three landmarks (actual location: 10, 20, and 30 cm) and a distal judgment of 60 cm for contact at the last three landmarks (actual location: 40, 50, and 60 cm). For all other experiments, this was modelled as a proximal judgment of 13 cm for contact at the first three landmarks (actual location: 13, 23, and 33 cm) and a distal judgment of 73 cm for contact at the last four landmarks (actual location: 43, 53, 63, and 73 cm). AIC was used to compare the fit of the sensory projection model with that of the sensory embodiment model (that is, the equality line in a linear regression; Fig. 1b, right inset). The model with the lower AIC score provides a better fit to the data. A significance cut off for the difference between fits (ΔAIC) was set to 3.22, which means that the better model is five times more likely to explain the pattern of judgments than the worse model.

In experiment 6, we sought to determine whether participants accurately rescaled their internal map of the rod to their estimated rigid:non-rigid ratio. All participants except one reported that the rod had both a rigid and non-rigid part and all participants over-estimated the actual contribution of the rigid portion to the overall length of the rod (0.76 ± 0.05). We hypothesized that a rescaling of this map would manifest itself as an over-estimation (that is, above the equality line) of where on the drawing they localized impacts on the rigid portion. Each participant's proportion estimate was used to normalize their localization judgments to compare the actual and ideal pattern of rescaling.

Analysis of vibrations. We first investigated whether impact at distinct locations on a tool led to highly reproducible vibration patterns, termed vibratory motifs. We measured the cross-correlation between every possible unique pairwise comparison for vibrations (time window: 0–100 ms) within each location (190 per location, 1,330 in total). This was done separately for each surface and for each participant. The distribution of values was characterized by taking the median cross-correlation value and the interquartile range.

We next classified impact location from vibratory motifs using a support vector machine[31] (SVM) with a radial basis kernel. Our classification scheme used fivefold cross validation. Thus, we trained the classifier on four subsamples of the data (that is, 80% of the trials) and tested classifier performance on the leftover subsample (that is, the remaining 20% of the trials). Each fold had an equal number of items per impact location. The hyperparameters of the SVM—C and γ—were tuned using grid search; tuning occurred separately for each of the five classification iterations. We specifically sought to characterize when a location-specific pattern emerged in the modal response of the rod. The features for classification were therefore subsets of the modal response across multiple temporal window sizes (2–60 ms, in steps of 2 ms). Classification was performed using the e1071 package (https://cran.r-project.org/web/packages/e1071/) and its interface with LIBSVM[32]. This was implemented with R version 3.2.3[33]. Chance classification (that is, random guessing) was ∼14%.

Analysis of afferent responses. Impact location was classified from the summed millisecond-precise spiking of the entire Pacinian (PC) population (0–100 ms after impact). As before, we used an SVM (fivefold cross validation) for classification with hyperparameters that were tuned using grid search. The features for classification were subsets of the spiking response across multiple time window sizes (5 to 100 ms, by steps of 5 ms). These results can be thought of as providing a theoretical lower bound for the accumulation of location information in the nervous system. Chance classification was ∼14%.

The coding of tactile information by the somatosensory system is often dependent upon the millisecond precise spiking of its first-order neurons[26,34–36]. To investigate whether impact location coding was dependent upon millisecond precision, we convolved the population-level spiking response on each trial with a Gaussian kernel at six distinct widths (1, 2, 4, 8, 16 and 32 ms). We then used an SVM to classify impact location from the first 50 ms of its response. This process was repeated for each of the six kernel widths and for each participant's dataset separately.

We further investigated whether trial-by-trial localization can be modelled as a function of the corresponding putative spiking response of the PC population (0–150 ms after impact) using multivariate regression. We initially focused exclusively on the spike-timing of the PC population (1 ms resolution). The FWDselect[37] package in R was used to select a range of time-points (10–50 predictor variables) that provided the best model fit as determined with Akaike information criterion (AIC; fivefold cross validation). R^2 and root-mean squared error were used to assess the fit of the final model. Leave-one-out cross-validation was then used to assess the predictive power of the model. The value for every trial was estimated with separate regression fit to the remaining 139 trials. We then calculated a predictive R^2 from the relationship between estimated and actual behaviour. This procedure was done separately for each participant.

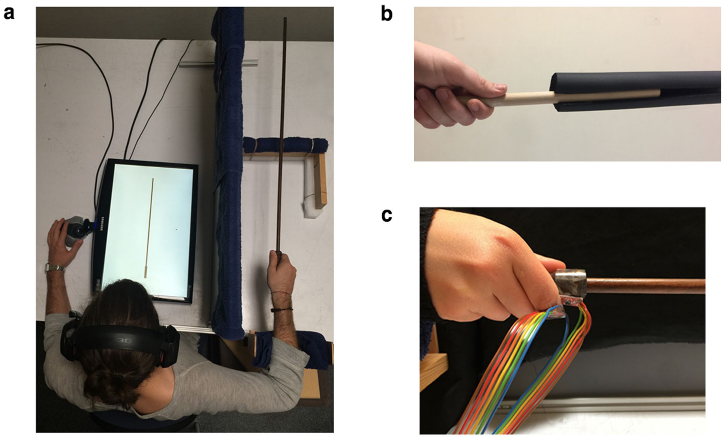
We next attempted to compare the ability of a spike-timing code and a rate code to model each participant's behaviour. We repeated the methods described above to model the trial-by-trial localization as a function of the firing rate of each of the 42 afferents. This was done for each participant separately. We compared the performance of the rate code and a spiking model (42 predictors) using AIC.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

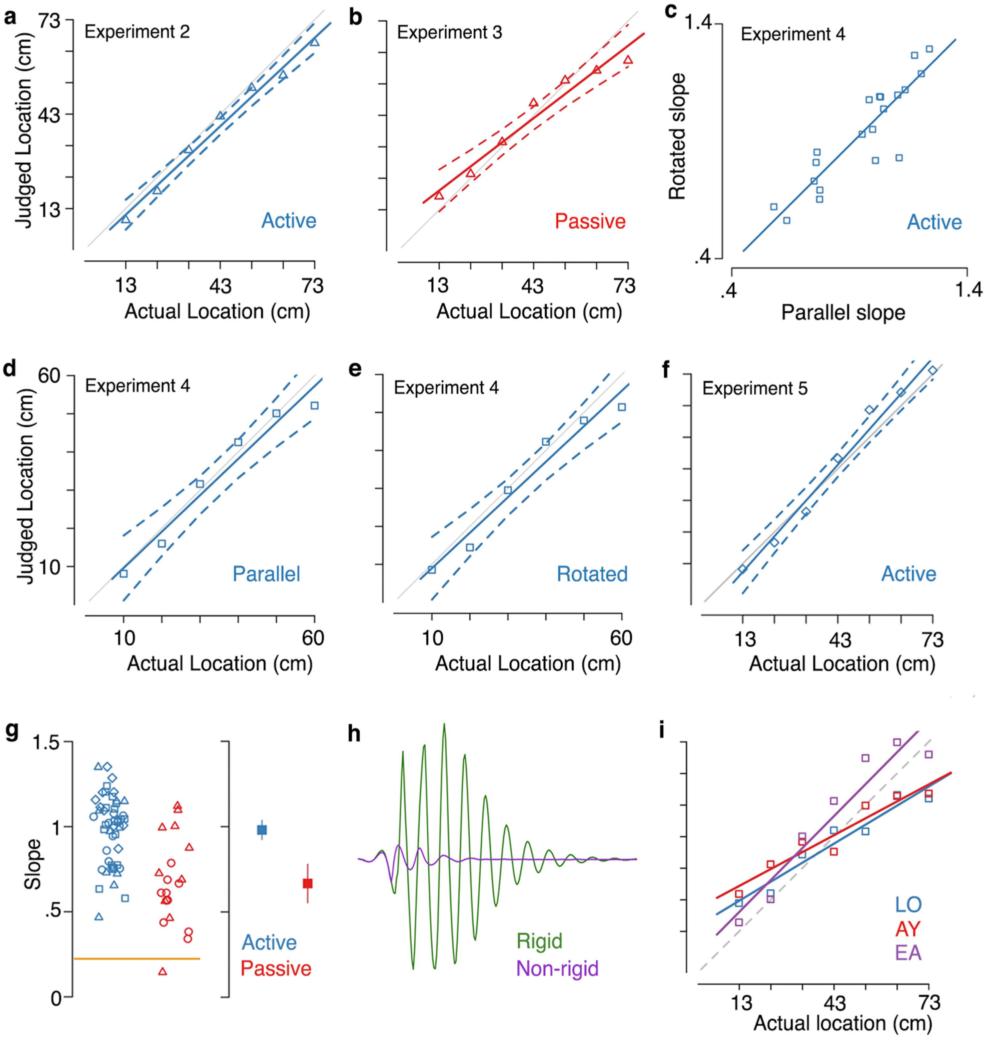
Code availability. Analysis code will be made available upon request to the corresponding authors, L.E.M. and/or A.F.

Data availability. All data has been archived at the Open Science Framework (https://osf.io/283cq/).

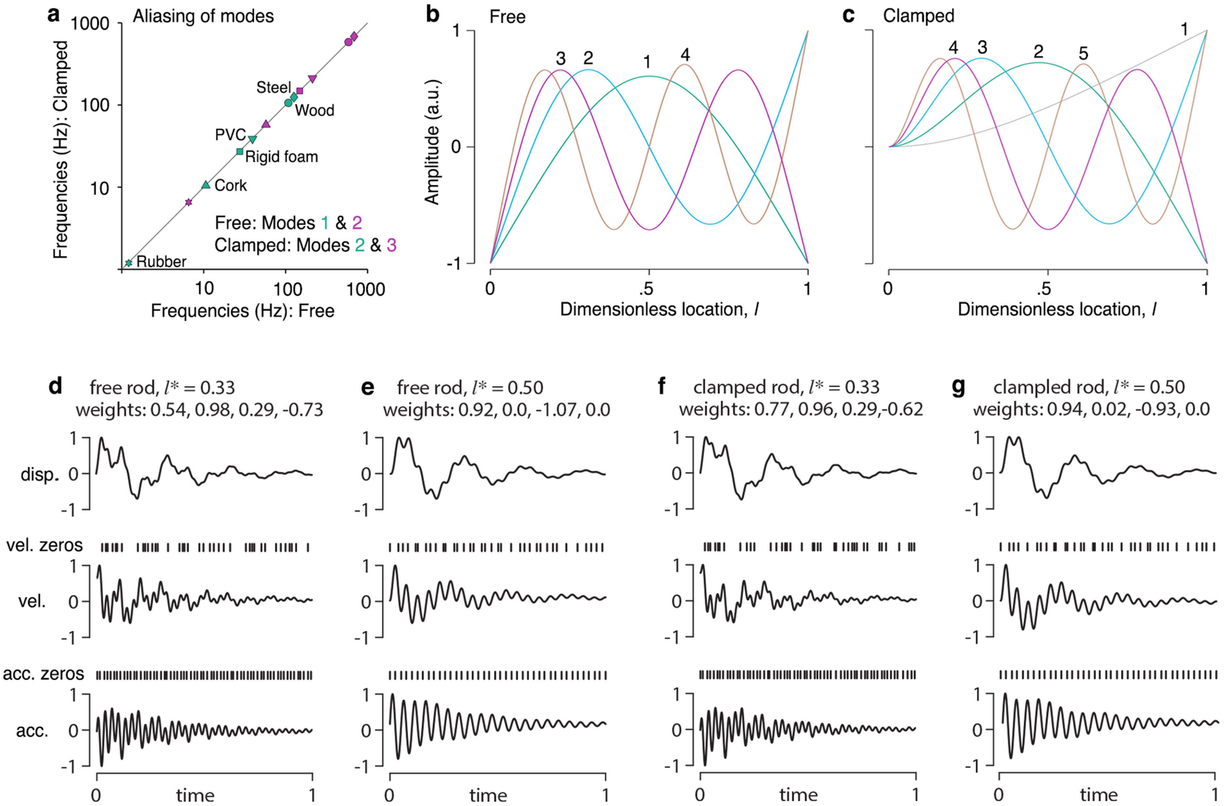
31. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20**, 273–297 (1995).
32. Chang, C.-C. & Lin, C.-J. LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2**, 27 (2011).
33. Team, R. C. R. *A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, 2014).
34. Pruszynski, J. A. & Johansson, R. S. Edge-orientation processing in first-order tactile neurons. *Nat. Neurosci.* **17**, 1404–1409 (2014).
35. Hayward, V. et al. Spatio-temporal skin strain distributions evoke low variability spike responses in cuneate neurons. *J. R. Soc. Interface* **11**, 20131015 (2014).
36. Jörntell, H. et al. Segregation of tactile input features in neurons of the cuneate nucleus. *Neuron* **83**, 1444–1452 (2014).
37. Sestelo, M., Villanueva, N. M., Meira-Machado, L. & Roca-Pardiñas, J. FWDselect: an R package for variable selection in regression models. *The R Journal* **8**, 132–148 (2016)



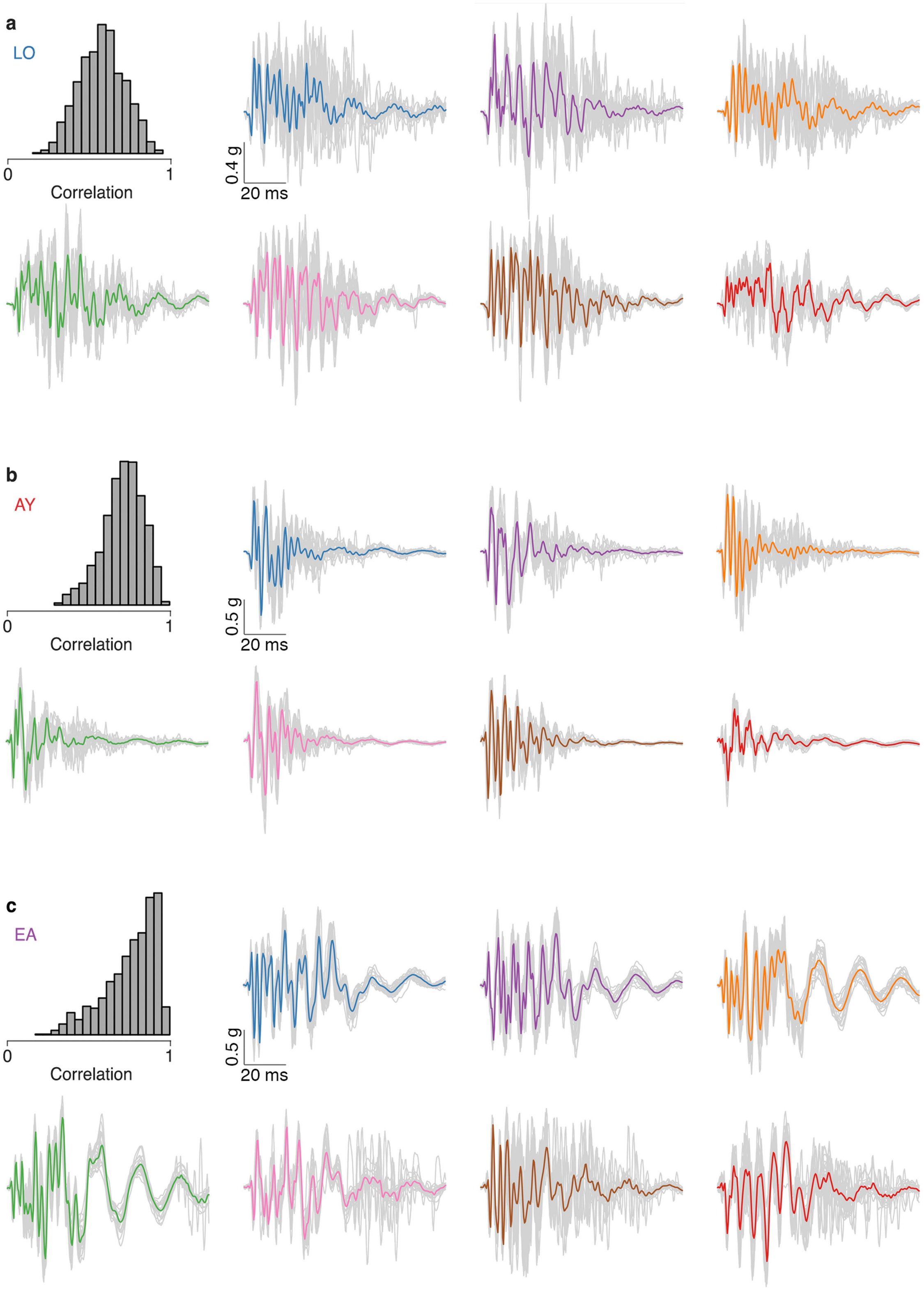
Extended Data Fig. 1 | Setup for the behavioural and vibration experiments. a, The experimental setup for all behavioural experiments (see Methods for details). The object shown below the rod was used in the first three behavioural experiments. **b**, The hybrid tool used in experiment 6, which was half rigid (wood and insulation) and half non-rigid (insulation only). The foam is displayed as being open for presentation purposes only. The dimensions of the rod have also been altered for presentation purposes. See Methods for more details. **c**, Setup for the vibration experiment, where accelerometers were attached to the handle and the participant's index finger.



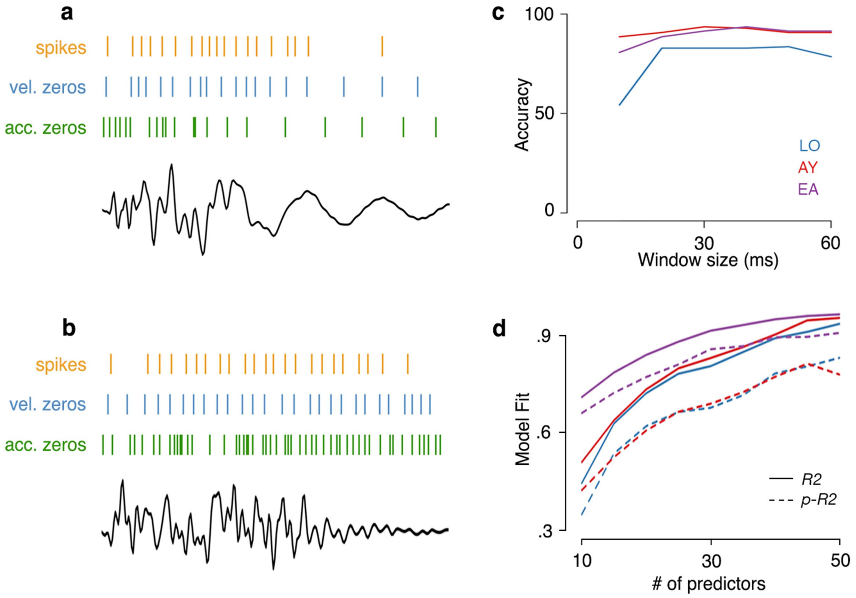
Extended Data Fig. 2 | Results of behavioural experiments. a–f, Group-level affine regression for experiments 2–5. Experiments 1, 2, 3 and 5 (*n* = 10), experiment 4 (*n* = 20). Coloured dashed lines around the model fits correspond to its 95% confidence interval. The grey line corresponds to the equality line. **c**, Experiment 4: Pearson's correlation between the regression slopes for when the drawing was displayed in parallel with the actual rod and rotated 90° counter-clockwise. **g**, Left, slope for every dataset from experiments 1–5 (*n* = 60). The prediction of the distalization model (that is, chance performance) is shown by the orange line. Right, average slope with 95% confidence intervals. **h**, Experiment 6, contact at identical locations on a wood (green) and foam (purple) tool leads to drastically different vibration patterns. In the case of the hybrid tool, participants could only feel the wooden portion of the rod with their hand, making the vibration pattern from the foam portion of the rod unexpected and therefore uninformative. **i**, Localization for each participant in the vibration experiment was within the range of behaviour observed in the other six experiments.



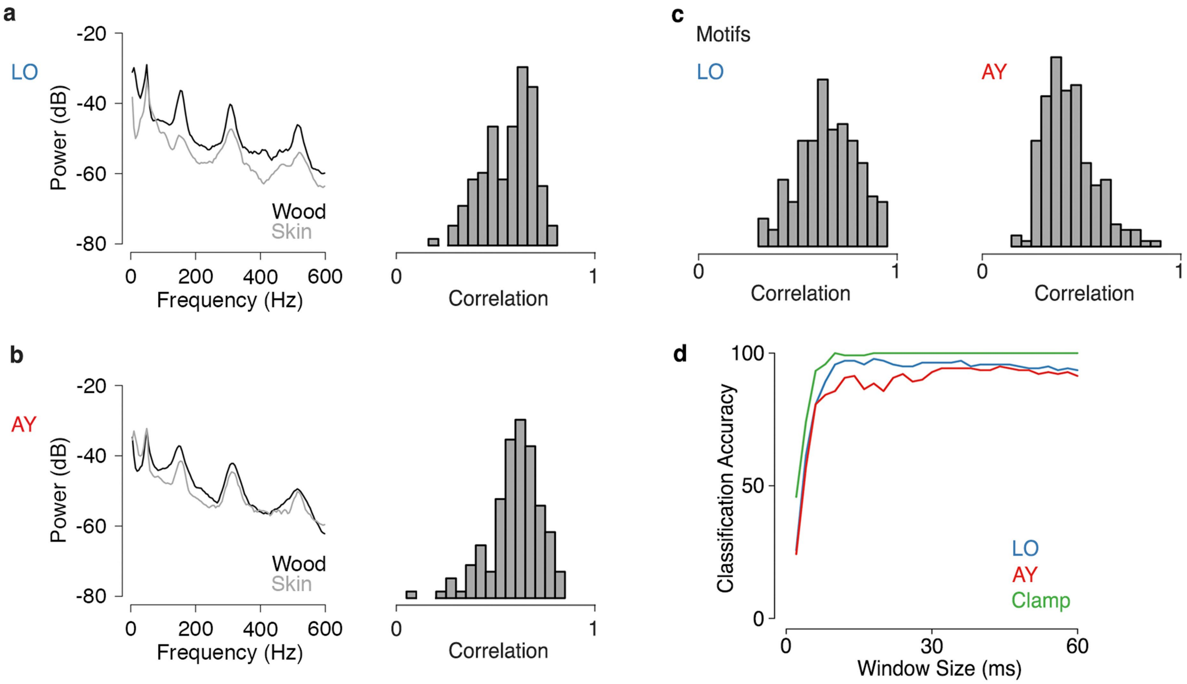
Extended Data Fig. 3 | Modal responses of a rod under different initial and boundary conditions. a, Independent of the material of a rod, the modal frequencies in the free case alias to the next highest modes in the clamped case. In the plot, this can be clearly seen for a simulated rod with a circular cross-section (length: 83 cm; cross-section radius: 0.8 cm) that was made of one of a diverse set of materials of varying elasticities and densities. The grey line corresponds to the equality line between the frequencies of each limit case. **b**, Mode shapes for the first four modes in the free case. **c**, Modes shapes for the first five modes in the clamped case. Modes with similar shapes as the free case are matched by colour. **d**, **e**, Simulated displacement, velocity and acceleration after impact on a free rod at $l^* = 0.33$ (**d**) and 0.5 (**e**). The zero crossings of velocity and acceleration are presented as tick marks above the relevant curves. The weights are taken from the mode shapes. **f**, **g**, Same for the clamped case, but minus the 'whipping' first mode. This is justified as all other mode shapes and frequencies are shared between cases. Notably, there is a high degree of similarity between the modal responses in each case for impacts at identical locations. Furthermore, it can be noticed in all panels that the sequences of zero crossings tend to repeat themselves owing to the special distribution of modal frequencies reflected in the phase differences. After a few periods of the low frequency modes these sequences can generally be easily discriminated. Thus, an effective feature space could be simply a relatively small number of time intervals between extrema in a suitably filtered signal.



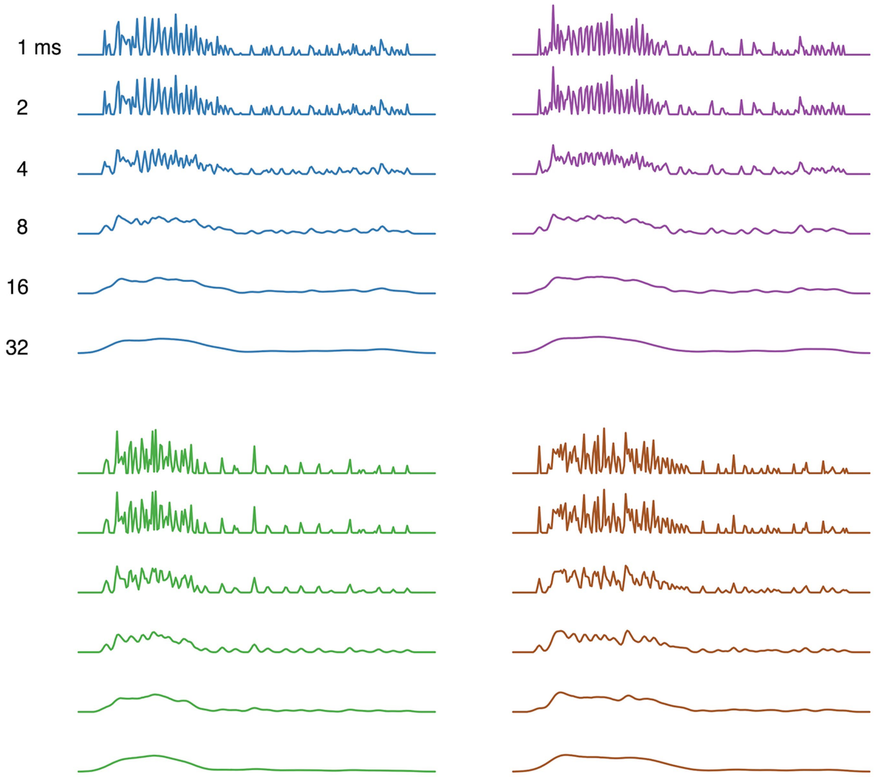
Extended Data Fig. 4 | Trial-by-trial vibrations for each participant were highly consistent. a–c, When held in the hand, vibrations after impact at each location on the tool were highly consistent for each participant: LO (**a**), AY (**b**) and EA (**c**). The upper left plot in each panel corresponds to the histogram of each within-location Pearson's correlation (0–100 ms after impact, corresponding to 250 data points per test). The shift in the distribution towards high correlations for each participant (LO: median $r = 0.58$, interquartile range (IQR) = 0.19; AY: median $r = 0.73$, IQR = 0.16; EA: median $r = 0.79$, IQR = 0.23) provides evidence for the emergence of vibratory motifs during passive and active sensing. The traces correspond to the motifs for each location (landmark 1 (blue) to landmark 7 (red), from left to right). The grey traces correspond to each individual impact and the colour traces correspond to the mean trace (0–100 ms after impact; colour coded by location).



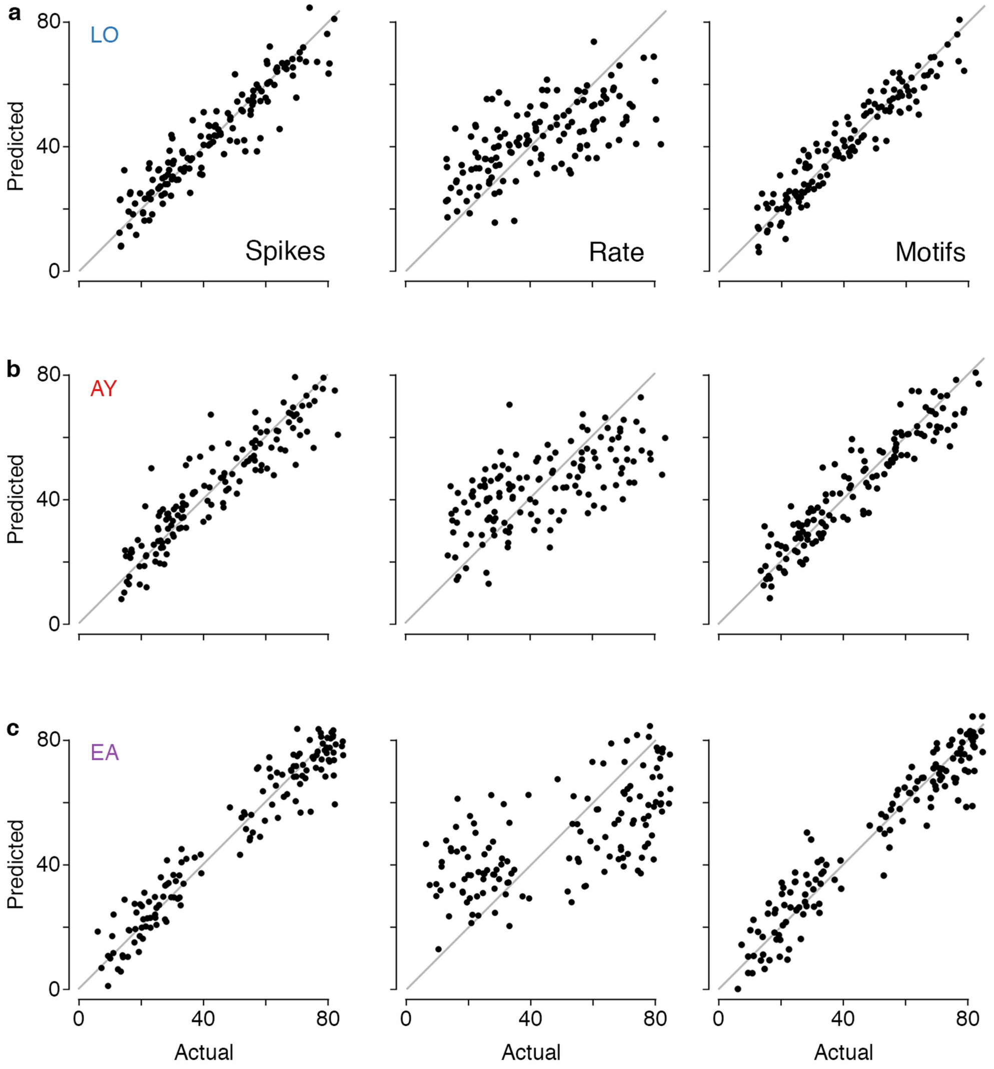
Extended Data Fig. 5 | The dimensionality of motifs can be reduced to their zero crossings. **a**, **b**, The motif (black; acceleration), zero crossing in acceleration (green) and velocity (blue), and spikes (orange) of a representative trial from the datasets of participants (**a**) EA and (**b**) LO. We observe a precise temporal relationship between zero crossings in velocity (blue) and the spikes of a Pacinian mechanoreceptor simulated with TouchSim (orange). **c**, We observed high classification accuracy when decoding impact location from the pattern of zero crossings in acceleration of motifs for each dataset (chance = approximately 14%). **d**, We could accurately model each participant's trial-by-trial behaviour given the temporal pattern of zero crossing in the acceleration of the corresponding motif (140 data points per test). We plot the goodness of fit as a function of the number of predictors used in the model. Solid lines correspond to the R^2 and dashed lines correspond to the predictive R^2.



Extended Data Fig. 6 | Results for skin recordings. a, **b**, We found similar spectral content between the mechanical (tool) and cutaneous (D2m) vibrations for both LO (**a**) and AY (**b**). Both participants showed similar peaks in the power spectrum for the vibrations (0–200 ms after impact) on the tool (black line) and skin (grey line). Furthermore, the trial-by-trial correlations between the spectral content of the wood and skin vibrations (histograms on the right) were high for both LO and AY. **c**, Histograms of all within-location comparisons for participants LO (left; median $r = 0.65$, IQR = 0.21) and AY (right; median $r = 0.41$, IQR = 0.17). Pearson's correlation on the first 100 ms after impact, corresponding to 250 data points per test. **d**, The observed speed that location information accumulated within the vibrations on the skin was extremely rapid for each participant, mirroring what was observed when the tool was clamped with a bench vice (green line; see the experiment in the Supplementary Data section 3).



Extended Data Fig. 7 | The effect of temporal smoothing on population spiking. Population-level spiking (−15–215 ms after impact) of the simulated afferents for randomly chosen trials from four different locations: landmark 1 (blue; trial 12), landmark 2 (purple; trial 39), landmark 4 (green; trial 68), and landmark 6 (brown; trial 114). To reduce the temporal resolution of spiking, we smoothed the response with Gaussian kernels at six different widths. The resulting traces were then used to investigate whether impact location coding was dependent on millisecond-resolution spike-timing (see Fig. 3c). Only trials from the dataset of participant EA are shown in this figure, but nearly identical patterns of results were found for both LO and AY.



Extended Data Fig. 8 | Model fits for each participant. **a**–**c**, The trial-by-trial population spike-timing of the putative afferents (left plots) precisely predicted the behaviour of each participant: LO (**a**), AY (**b**), and EA (**c**). When the population rate code (centre plots) was used as features, the model did not provide a precise fit to the behaviour (see Extended Data Table 2). Vibratory motifs (right plots) also predicted behaviour with high precision. The grey line in all plots represents the equality line between actual and predicted behaviour, not the actual regression line.

Extended Table 1 | Behavioural results

Exp.	*Condition*	*Slope*	*Intercept*	*R^2*	*t-Test Type*	*Comparison*	*t*	*P*
*1**	Active	.93±.09	-2.20±2.01	.94±.02	One-sample	Slope vs. 0.25	7.82	<.001
						Slope vs. 1	0.77	.46
						Intercept vs. 0	1.09	.33
	Passive	.57±.04	14.18±3.3	.85±.02	One-sample	Slope vs. 0.25	7.08	<.001
						Slope vs. 1	9.75	<.001
						Intercept vs. 0	4.30	.002
					Paired	Slope	5.43	<.001
						Intercept	4.63	.001
*2**	Active	.94±.04	-1.13±2.72	.93±.01	One-sample	Slope vs. 0.25	16.04	<.001
						Slope vs. 1	1.48	.17
						Intercept vs. 0	0.41	.69
*3**	Passive	.77±.1	9.11±3.76	.87±.04	One-sample	Slope vs. 0.25	5.22	<.001
						Slope vs. 1	2.34	.04
						Intercept vs. 0	2.43	.04
*4***	Parallel	.95±.04	0.08±1.72	.91±.01	One-sample	Slope vs. 0.25	16.08	<.001
						Slope vs. 1	1.08	.29
						Intercept vs. 0	0.04	.97
	Rotated	.94±.05	-0.39±1.78	.92±.01	One-sample	Slope vs. 0.25	14.04	<.001
						Slope vs. 1	1.32	.20
						Intercept vs. 0	0.22	.83
					Paired	Slope	0.88	.44
						Intercept	0.53	.61
*5**	Active	1.13±.05	-4.30±3.38	.95±.01	One-sample	Slope vs. 0.25	19.40	<.001
						Slope vs. 1	2.84	.02
						Intercept vs. 0	1.27	.24

Statistical tests were two-sided. Summary values are presented as mean ± s.e.m.

**n* = 10.

***n* = 20.

Extended Table 2 | Multivariate models with forty-two predictor variables

Participant[%]	*Sensing*	*Features*	$F_{42,97}$	R^2	p-R^2	*RMSE*	*AIC**
LO	Passive	Spike timing	11.04	0.83	0.65	7.20	1038.07**
		Rate	1.70	0.42	-0.22	13.13	1206.25
		Motifs	22.52	0.91	0.83	4.72	
AY	Passive	Spike timing	10.02	0.82	0.65	8.17	1073.46**
		Rate	1.62	0.45	-0.20	14.05	1225.16
		Motifs	17.9	0.89	0.77	6.38	
EA	Active	Spike timing	20.37	0.90	0.79	8.14	1072.55**
		Rate	1.73	0.37	-0.23	20.21	1327.09
		Motifs	26.85	0.92	0.85	7.18	

p-R^2, predictive R^2; RMSE, root-mean squared error.

[%] 140 data points per regression.

*44 degrees of freedom per model.

**Significantly better model.

LETTER

https://doi.org/10.1038/s41586-018-0477-4

In vivo reprogramming of wound-resident cells generates skin epithelial tissue

Masakazu Kurita[1,2], Toshikazu Araoka[1,3], Tomoaki Hishida[1], David D. O'Keefe[1], Yuta Takahashi[1], Akihisa Sakamoto[1,3], Masahiro Sakurai[1,3], Keiichiro Suzuki[1], Jun Wu[1], Mako Yamamoto[1], Reyna Hernandez-Benitez[1], Alejandro Ocampo[1], Pradeep Reddy[1], Maxim Nikolaievich Shokhirev[4], Pierre Magistretti[5], Estrella Núñez Delicado[3], Hitomi Eto[2], Kiyonori Harii[2] & Juan Carlos Izpisua Belmonte[1]*

Large cutaneous ulcers are, in severe cases, life threatening[1,2]. As the global population ages, non-healing ulcers are becoming increasingly common[1,2]. Treatment currently requires the transplantation of pre-existing epithelial components, such as skin grafts, or therapy using cultured cells[2]. Here we develop alternative supplies of epidermal coverage for the treatment of these kinds of wounds. We generated expandable epithelial tissues using in vivo reprogramming of wound-resident mesenchymal cells. Transduction of four transcription factors that specify the skin-cell lineage enabled efficient and rapid de novo epithelialization from the surface of cutaneous ulcers in mice. Our findings may provide a new therapeutic avenue for treating skin wounds and could be extended to other disease situations in which tissue homeostasis and repair are impaired.

The epidermis is the outermost layer of the body, helping to maintain organismal homeostasis by protecting against environmental insults and preventing water loss. The multi-layered epidermis is maintained by stem and progenitor cells within the basal layer[1] (that is, basal keratinocytes). An important step in repairing damaged skin is the migration of keratinocytes from adjacent epidermis into the wound to promote re-epithelialization[3]. For large wounds, this process is inefficient. To improve patient outcomes, more rapid and efficient methods for regenerating epidermal coverage must be developed, preferably non-surgical interventions. After recent advances in cellular reprogramming[4,5], we reasoned that wound-resident cells could be reprogrammed towards an epidermal progenitor cell fate to generate new epithelial cells, thereby promoting de novo epithelialization from the surface of a cutaneous ulcer.

We initially sought to reprogram mesenchymal cells (human dermal fibroblasts (hDFs) and adipose-derived stromal cells (hADSCs)), as they participate in wound healing[3,6]. By comparing gene-expression profiles of human keratinocytes and primary hDFs (Extended Data Fig. 1a, b), upstream-promoter analyses (Supplementary Tables 1–3) and gene-expression reversal analysis (Extended Data Fig. 1c, Supplementary Table 4), we identified 55 transcription factors and 31 microRNAs that are potentially involved in keratinocyte specification (Supplementary Table 5). These candidates were further tested by assessing expression levels after calcium-induced differentiation of keratinocytes (Extended Data Fig. 1d), and by transducing each candidate into hDFs and measuring the levels of the keratinocyte markers, keratin 14 (KRT14) and cadherin 1 (Extended Data Fig. 1e). Selected factors were then co-transduced (via lentivirus) in different combinations into hDFs and the generation of keratinocyte-like cells was assessed (Extended Data Fig. 1f–k, Supplementary Table 6). A combination of 28 transcription factors generated keratinocyte-like cells in vitro. When these cells were grown in 3D organotypic culture, they formed a stratified epithelium (Extended Data Fig. 1l). We therefore named these cells 28TF-induced stratified epithelial progenitors (28TF-iSEPs).

To exclude the possibility that keratinocytes were contaminating the cell-isolation process, we switched to working with hADSCs. To improve transduction efficiency and lower cytotoxicity, we switched to retroviruses (Extended Data Fig. 1n, o). Using this system, we systematically eliminated redundant factors by: (1) excluding factors not integrated into the genome of 28TF-iSEPs (Extended Data Fig. 1m); (2) removing one factor at a time[4]; and (3) measuring keratinocyte

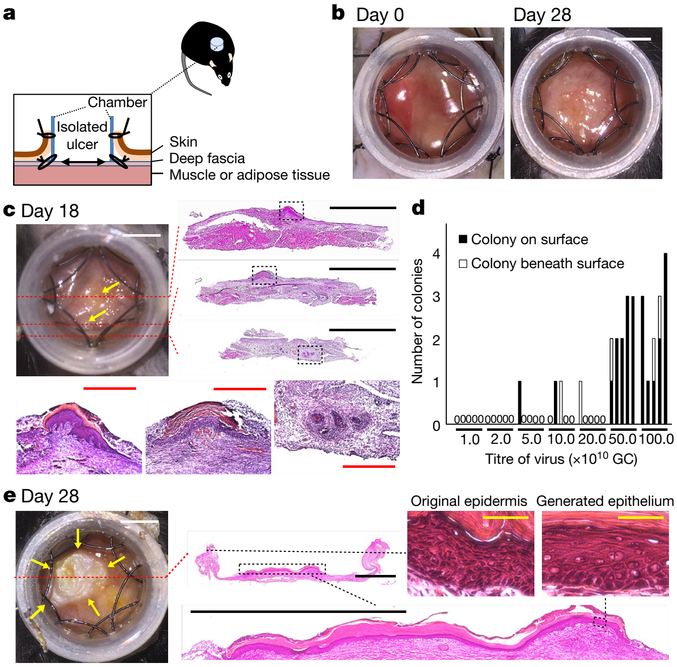


Fig. 1 | DGTM factors generate epithelial tissues through in vivo reprogramming. a, Schematic of the skin chamber used for separating the ulcer from surrounding skin. **b**, An ulcer in the chamber on the day of attachment (left) and 28 days later (right). Scale bars, 3 mm. $n = 20$, all similar results. **c**, Chambered ulcer 18 days after administration of DGTM-AAV (1.0×10^{12} gene copies). Haematoxylin and eosin (H&E) staining of sections through the generated epithelium. Red dotted lines indicate locations of the sections; yellow arrows indicate the generated epithelium. Black and white scale bars, 3 mm; red scale bars, 500 µm. $n = 5$, all similar results. **d**, Number of epithelial colonies 18 days after DGTM-AAV administration (titre indicated). GC, gene copies. **e**, Left, representative image showing an ulcer 28 days after DGTM-AAV administration. Yellow arrows indicate the periphery of the generated epithelium. Red dotted line indicates the position of the section shown middle top. Right (bottom), H&E staining of the generated epithelium. Right (top), magnified panels showing H&E staining of original skin and the generated epithelium. Black and white scale bars, 3 mm; yellow scale bars, 50 µm. $n = 21$, all similar results.

---

[1]The Salk Institute for Biological Studies, La Jolla, CA, USA. [2]Department of Plastic Surgery, Kyorin University School of Medicine, Tokyo, Japan. [3]Universidad Católica San Antonio de Murcia (UCAM), Campus de los Jerónimos, Guadalupe, Spain. [4]The Razavi Newman Integrative Genomics & Bioinformatics Core, The Salk Institute for Biological Studies, La Jolla, CA, USA. [5]King Abdullah University of Science & Technology (KAUST), Thuwal, Saudi Arabia. *e-mail: belmonte@salk.edu

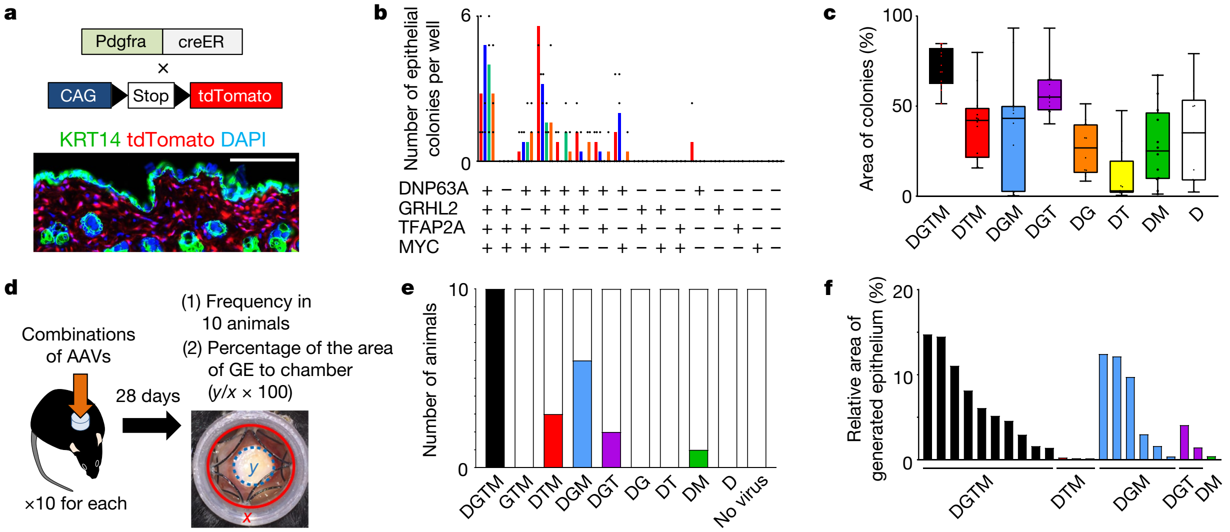


Fig. 2 | DGTM factors work collaboratively to generate epithelial tissues. a, Lineage tracing using *Pdgfra*^{creER};*LSL*^{tdTomato} mice. A representative image showing tdTomato fluorescence and KRT14 localization in skin from the back of one such mouse. Scale bar, 200 μm. Findings were confirmed in 42 animals used for lineage tracing studies. **b**, Number of epithelial shaped colonies 14 days after indicated combinations of transcription factors were transduced in vitro. Four cell lines from different animals were tested. Each cell line is represented by a different colour ($n = 3$ technical replicates). Overlaid dot plot indicates the distribution of the data. **c**, Surface areas of epithelial colonies derived from single cells. Transcription factor combinations are indicated. Whiskers represent the maximum and minimum, the box represents the range from 25th to 75th percentile and the centre line is the median. Overlaid dot plot indicates the distribution of the data ($n = 11$ for D, $n = 12$ for the rest). **d**, Experimental design schematic with a representative image of a DGM-AAV-treated wound. Red and dotted blue lines indicate area of the chambered ulcer (*x*) and the area of the generated epithelium (GE) (*y*), respectively. **e**, Number of animals with (colour bars) or without (white bars) generated epithelial tissues. **f**, Surface area of generated epithelial tissues.

phenotypes (Supplementary Table 7). Transduction of *DNP63A* (deltaNp63alpha, an isoform of *TP63*) and *GRHL2* reprogrammed hADSCs into cells similar to 28TF-iSEPs (Extended Data Fig. 1p–s). An additional round of screening revealed that: (1) *MYC* (also known as *c-MYC*) enhanced reprogramming efficiency, cell proliferation and epithelial stratification, and (2) *TFAP2A* quickened the emergence of colonies (Supplementary Table 8, Extended Data Fig. 2a–m). Thus, the optimal combination of reprogramming factors was *DNP63A*, *GRHL2*, *TFAP2A* and *MYC* (DGTM factors). It is important to note that these in vitro-generated epithelia lacked cornification and expressed keratin 13 in the suprabasal layer (characteristic of mucosal epithelium[7], foreskin epidermis[8], and hyper-proliferative skin[9]), not keratin 10 (characteristic of adult skin epidermis[7]) (Extended Data Fig. 2h, m). As epithelia are potentially influenced by their niche[10], we hypothesized that applying DGTM factors in vivo would more effectively generate skin-like epithelial tissue.

To investigate whether in vivo reprogramming via DGTM factors could induce de novo generation of epithelial tissue from the surface of a skin ulcer, we developed an assay of isolated skin ulcers that simulated the central portion of a large cutaneous ulcer. We surgically removed skin from the back of mice to generate an ulcer and isolated the resulting wound from the surrounding skin using a skin chamber sutured to the deep fascia. This serves to prevent migration of keratinocytes into the wound, and closure of the wound by contraction (Fig. 1a). The absence of epithelial cells within these wounds was confirmed using *Krt14*^{cre};Rosa-CAG-loxP-stop-loxP (LSL)-tdTomato (hereafter *Krt14*^{cre};*LSL*^{tdTomato}) mice, in which all cells that express *Krt14* sometime during their lifetime are labelled (Extended Data Fig. 3a–h). In the absence of treatment, isolated wounds did not re-epithelialize (Fig. 1b). To test the DGTM factors in vivo, we switched to adeno-associated viruses (AAVs). As AAV cell tropisms differ between serotypes, we determined which AAV could be effective by subcutaneously injecting or directly applying green fluorescent protein (GFP)-expressing AAVs (GFP-AAV) to isolated ulcers. AAVDJ (a serotype of AAV capsid generated by DNA family shuffling[11]) resulted in the highest levels of GFP (Extended Data Fig. 3i–l). Furthermore, we assessed AAVDJ tissue organ distribution via injection of luciferase-AAVDJs into tail veins (that is, systemically), subcutaneous injection, or inoculation into the wound chamber. Regardless of the delivery method, luciferase expression distal to the injections site was primarily confined to the liver (Extended Data Fig. 3m).

We next administered DGTM-AAVs (DNP63A-AAV, GRHL2-AAV, TFAP2A-AAV, and MYC-AAV) in our in vivo ulcer assay. Within 18 days, we observed epithelia-like tissue inside the chamber (Fig. 1c). Histological analysis of day-18 samples resulting from different titres of AAV (Fig. 1d, Extended Data Fig. 4a), revealed that 5.0×10^{11} gene copies of each factor was the minimal titre needed to efficiently generate epithelial tissue. This titre was used for subsequent in vivo experiments. Twenty-eight days after transducing isolated wounds with DGTM-AAVs, de novo epithelial tissues were histologically very similar to the skin adjacent to the wound edge (Fig. 1e).

To determine the proportion of transduced cells contributing to the healing process, we repeated these experiments using *Pdgfra*^{cre};*R26R*^{confetti} mice[12], in which mesenchymal cells are randomly labelled with GFP, YFP, or RFP, although not all mesenchymal cells were labelled. Each epithelial cluster was derived from a single mesenchymal cell and as these individual clones grew, they intermingled with one another to form a single epithelium. The efficiency of in vivo reprogramming was estimated to be approximately 0.1% (Extended Data Fig. 4b–i).

To assess the role of each DGTM factor in reprogramming mesenchymal cells to epithelial cells, we tested different combinations of DGTM AAVs (for example, D, G, T, but not M). First, a lineage tracing system was introduced using *Pdgfra*^{creER};*LSL*^{tdTomato} mice[6]. Preoperative administration of tamoxifen labelled a broad spectrum of mesenchymal cells (but not epithelial cells) with tdTomato (Fig. 2a). Different combinations of DGTM AAVs were tested in vitro using Pdgfra^{+} mouse ADSCs (mADSCs) sorted from *Pdgfra*^{creER};*LSL*^{tdTomato} mice on the basis of tdTomato fluorescence (Extended Data Fig. 5a–c). After confirmation of in vitro reprogramming of mouse mesenchymal cells towards iSEPs with DGTM-AAVs (Extended Data Fig. 5d–j), different combinations of DGTM factors were tested at relatively lower titre, and the number of epithelial-like colonies counted on day 14 (Extended Data Fig. 5k–n). Epithelial-like colonies arose for all combinations

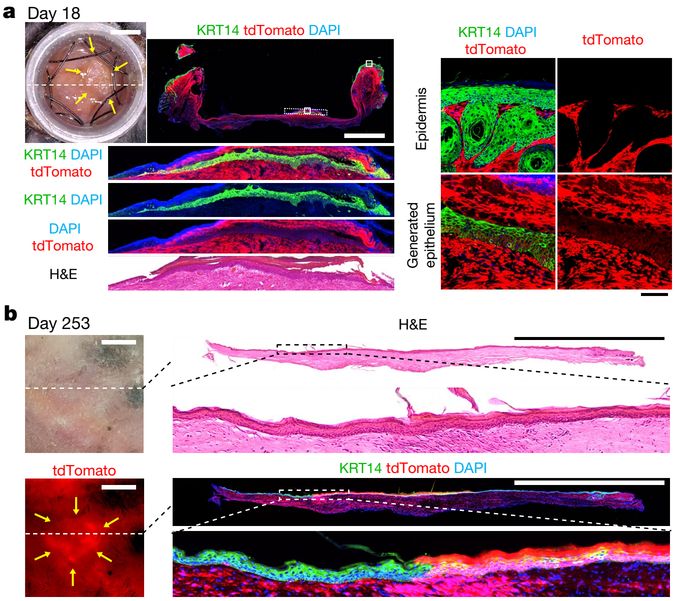


Fig. 3 | Generated epithelial tissue enables wound healing. a, Appearance (top left) and immunohistochemical analysis (top middle) of an ulcer 18 days after DGTM-AAV administration in *Pdgfra*^{creER};*LSL*^{tdTomato} mice. Localization of KRT14 and tdTomato are shown. In the photograph, the yellow arrows indicate the periphery of the generated epithelium and the white dotted line indicates the position of the histological section. In the immunohistochemistry image the white dotted outline indicates the position of magnified panels (bottom) and the white solid outline indicates the position of the highly magnified views of the epidermis and generated epithelium (right). White scale bar, 3 mm; black scale bar, 100 μm. Images are representative of the three independent experiments. **b**, Appearance (top left), stereoscopic analysis (bottom left), H&E staining (top right), and immunohistochemical analysis (bottom right) of skin and subcutaneous tissue including generated epithelium on day 253. Dotted lines in the left panels indicate the position of the sections shown on the right. Stereoscopic analysis revealed the area of generated tissue that is not evident by appearance. Dotted outlines indicate the positions of magnified panels. Scale bars, 5 mm. Similar findings were confirmed in all animals tracked for 6–9 months ($n = 5$).

that contained *DNP63A* (Fig. 2b). All *DNP63A*-containing combinations resulted in cells similar to primary keratinocytes (Extended Data Fig. 5o, p). Cumulative evidence suggests that the clonogenicity of cultured keratinocytes reflects their stemness, and thus their potential for repairing and maintaining epidermal tissue[13,14]. We therefore assessed the clonogenicity of epithelial cells generated using each of the *DNP63A*-containing combinations. On the basis of the expansion of single-cell clones (Extended Data Fig. 5q), clonogenicity was enhanced as the number of transduced factors increased (Fig. 2c). Next, we analysed the efficacy of the eight *DNP63A*-containing combinations (as well as GTM) in generating epithelial tissue in vivo using isolated skin ulcers, analysing the frequency and size of generated epithelia on day 28. These analyses also suggested an indispensable role for *DNP63A* in reprogramming (Fig. 2d–f, Extended Data Fig. 5r, s). For both frequency and size of the generated epithelia, GTM factors worked collaboratively with *DNP63A* to generate epithelial tissues. Notably, the oncogene *MYC* was dispensable for the de novo generation of epithelial tissues. Considering the in vitro emergence of epithelial cells and clonal ability, in vivo ulcer conditions (such as inflammation) may further constrain the generation and survival of epithelial tissue, thus leaving a relatively limited number of factor combinations that can effectively generate epithelial tissues in vivo.

To determine whether DGTM-generated epithelia are maintained over extended periods of time and whether it could promote the healing of large wounds, we subjected mice to multiple rounds of wounding. We used *Pdgfra*^{creER};*LSL*^{tdTomato} mice for these tests to ensure that DGTM-generated epithelia originated solely from mesenchymal cells (Fig. 3a, Extended Data Fig. 6a–e). After the generation of epithelial tissue within the isolated ulcer, the skin chamber was replaced with a larger one. Generated tissues laterally expanded within the large chamber (Extended Data Fig. 6f–h). After the large chamber was removed, generated tissue successfully connected to the surrounding epidermis while retaining a stratified epithelial structure (Fig. 3b, Extended Data Fig. 6i). As AAVDJ showed a specific tropism towards liver, we investigated possible histological alterations of the liver, complete blood count, and blood chemistry 3 months after AAV administration, with no evidence of pathology (Extended Data Fig. 6j–l).

To further assess the regenerative potential of DGTM-generated epithelial tissue, we generated ulcers within a large chamber, but left a small patch of skin intact in the centre of the ulcer. The ability of this skin island to expand and heal the ulcer could then be tested. We tested four types of islands: (1) intact skin, (2) epithelialized skin, (3) cell sheets and (4) DGTM-generated epithelia (Extended Data Fig. 6m–q). On average, epithelialization kinetics of intact skin, epithelialized skin, and DGTM-generated epithelia were similar, whereas the cell sheet group exhibited early contraction of the skin island and delayed epithelialization (Extended Data Fig. 6r, s).

To compare the histological properties of the generated epithelia and differentiated epidermis, samples were collected at day 18, days 28–30 and days 90–110 and analysed via immunohistochemistry. Similar to normal skin, epithelial tissues generated in vivo formed a cornifying envelope and expressed loricrin in the most superficial layers (Fig. 4a). Regarding keratins expressed in the suprabasal layer, generated epithelia could be classified into three types: keratin-10 (K10)$^{+}$keratin-13 (K13)$^{-}$, K10^{+}K13^{+} and K10^{-}K13^{+} (Fig. 4a, Extended Data Fig. 6t). K10 is expressed in the natural skin epidermis, whereas K13 is expressed in mucosal epithelium[7] (Fig. 4a) and hyperproliferative epidermis (for example, fetal skin[9]) (Extended Data Fig. 6u). Epithelial tissue that first emerged had sporadic K13 expression with or without K10, but this progressively shifted towards a K13^{-}K10^{+} pattern, thereby moving closer to the keratin pattern of natural skin epidermis (Fig. 4b). Thus, if allowed to mature for a sufficient amount of time within the wound niche, DGTM-generated epithelia exhibited histological characteristics of normal skin.

One of the most important functions of the skin is to serve as a barrier against environmental insults. To assess the outside–in barrier function of DGTM-generated epithelia, a toluidine blue dye penetration assay[15] was performed. On day 28, samples including the ulcer, DGTM-generated epithelia and surrounding skin were dissected and their external surfaces were exposed to toluidine blue dye. Similar to the surrounding skin, generated epithelial tissue effectively blocked dye penetration (Fig. 4c). To assess dye penetration after complete epithelialization, we used day-40 *Pdgfra*^{creER};*LSL*^{tdTomato} mice in which the skin chamber was removed on day 30. The generated epithelia and surrounding skin exhibited similar abilities to block penetration of lucifer yellow (after 1 h immersion)[15] (Fig. 4d). To measure inside–out barrier function, transepidermal water loss (TEWL)[15] was compared between generated epithelium (days 90–110), intact skin and a freshly created ulcer on the same animal. DGTM-generated tissues showed TEWL values equivalent to those of intact skin (Fig. 4e).

To assess the clinical relevance of this technique, we characterized the effect of DGTM-AAV on vascularity of the wound bed in our chamber model. DGTM-AAV showed no significant influence on vascularity of the wound bed (Extended Data Fig. 7a–d). Next, we investigated whether in vivo reprogramming could be used to heal an older wound. DGTM-AAVs were applied 7 days after creating the ulcer, and generation of an epithelium was observed in all ten animals tested (18 days after DGTM-AAV application). Notably, generated epithelia tended to be larger than seen with fresh ulcers (compare Extended Data Fig. 7e–g and Extended Data Fig. 4a). To improve the clinical relevance of this potential therapy, we sought to enhance the AAV delivery by administering AAVs via collagen gel. A GFP-AAV solution was mixed with an equal amount of collagen gel and then applied to ulcers in a chamber.

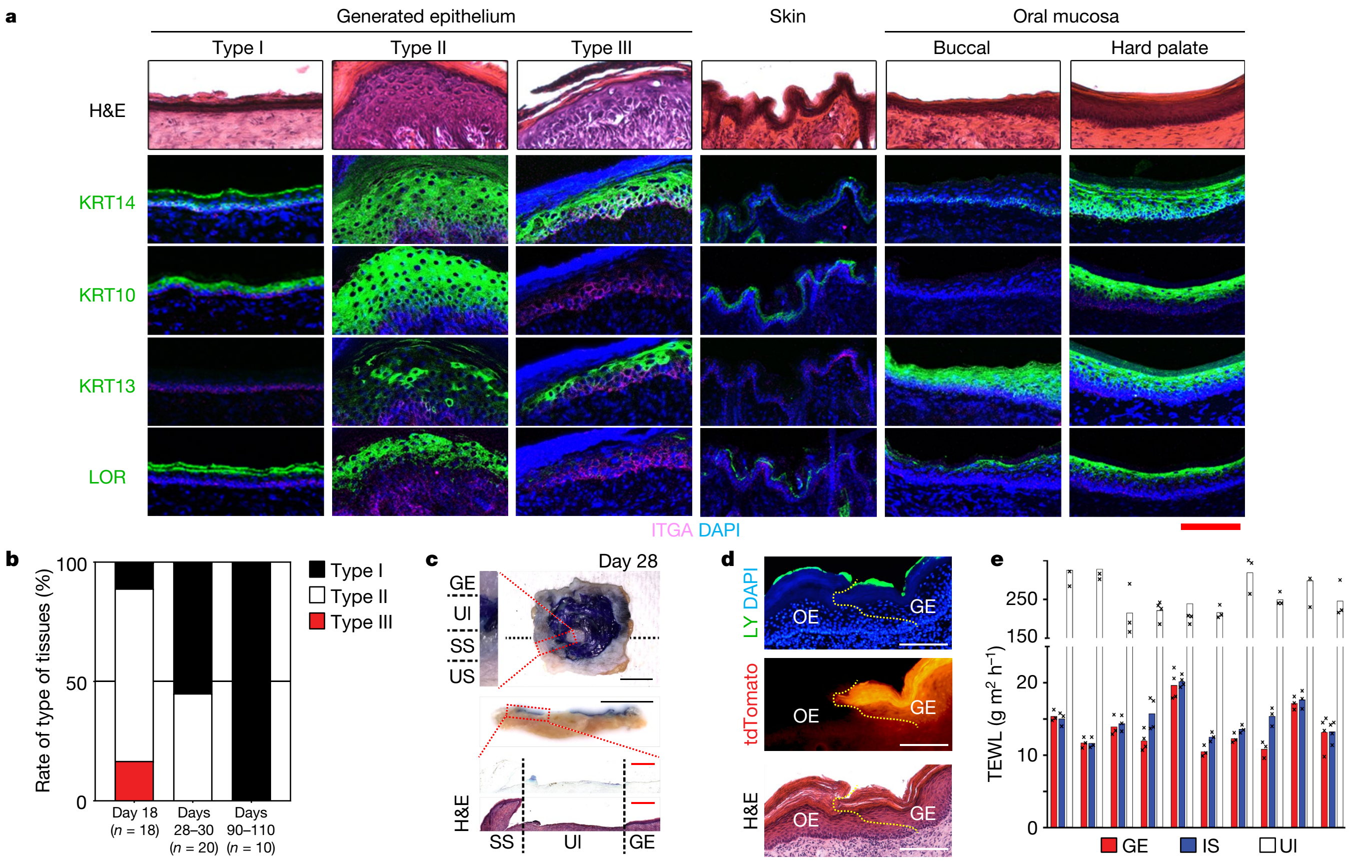


Fig. 4 | Histological characterization and functional barrier properties of generated epithelia. a, H&E staining and immunohistochemical analysis of different types of generated epithelium, as well as adult skin and oral mucosa. Generated epithelia were obtained at day 104 for Type I ($K10^+K13^-$) and day 18 for Types II ($K10^+K13^+$) and III ($K10^-K13^+$). Scale bar, 100 μm. Images of skin and oral mucosa are representative of three independent experiments. **b**, Percentage of Types I–III generated epithelia on day 18, days 20–30 and days 90–110. **c**, Top, toluidine blue staining assay with a magnified image of the outlined region shown to the right. Middle, cross-section of stained sample with magnified images of the outlined region shown below including toluidine blue and H&E staining. GE, generated epithelia; SS, stained skin; Ul, ulcer; US, unstained skin. Black scale bars, 5 mm; red scale bars, 500 μm. **d**, Lucifer yellow (LY) dye penetration assay. Middle and bottom panels are fluorescence and H&E staining images, respectively. These helped to identify the border between the original epidermis (OE) and generated epithelium. Scale bars, 200 μm. **c**, **d**, Images are representative of three animals. **e**, TEWL values of generated epithelium, intact skin (IS), and ulcer from ten animals (days 90–110). The overlaid dot plot indicates the distribution of the data. Data are from three or four technical replicates.

The collagen gel increased both GFP expression and residual AAV copy number in ulcer tissue, and reduced AAV copy number in the liver (Extended Data Fig. 7h–k). Thus, collagen gel increased both the potency and specificity of the AAV system.

We reasoned that DGTM-AAV effectiveness could be further improved by adding bioactive molecules, such as FGF2[16,17] and/or a Rock inhibitor[18]. To test this, DGTM-AAVs in a collagen gel were applied to an isolated wound. Subsequently, from day 4, wounds were treated with Rock inhibitor and/or FGF2 until day 17. When applied alone, Rock inhibitor and FGF2 both enhanced healing. For wounds treated with both Rock inhibitor and FGF2, the wound surface was almost completely epithelialized within 2–2.5 weeks (Extended Data Fig. 7l–m). Thus, these two factors, in combination with a collagen scaffold, greatly enhanced in vivo reprogramming efficiency, resulting in concomitant de novo epithelialization from different regions within the ulcer leading ultimately to rapid wound healing.

To characterize in vivo reprogrammed cells, iSEPs were sorted from generated epithelia on the basis of tdTomato fluorescence 1, 3 and 6 months after AAV-DGTM administration. To investigate stemness[13,14], clonogenic abilities of 3- and 6-month in vivo iSEPs were compared with primary keratinocytes from adult animals and one-day-old pups. In vivo iSEPs consistently showed higher clonogenic abilities than primary keratinocytes (Extended Data Fig. 8a–d). To determine whether this high clonogenic potential resulted from malignant transformation, we cultured in vivo iSEPs in soft agar and confirmed that they lacked anchorage-independent proliferative capacity (Extended Data Fig. 8e–f). To reveal their kinetics in vivo, 3-month in vivo iSEPs were subcutaneously transplanted into immunodeficient mice. For comparison, we also transplanted mouse embryonic stem cells (mES cells) that ubiquitously expressed GFP, HeLa cells that could be detected with a human-specific antibody, and mouse primary keratinocytes from $KRT14^{cre};LSL^{tdTomato}$ mice. In vivo iSEPs formed smooth, soft, round nodules that were larger than those formed by primary keratinocytes. Histologically, in vivo iSEPs nodules consisted of a large epithelial cyst and a small number of scattered cells within the transplant capsule, consistent with the histological characteristic of primary keratinocytes. Teratogenecity, as observed for mES cells, and transfascial invasive properties, as observed for HeLa cells, were not detected for in vivo iSEPs (Extended Data Fig. 8g–n). As a final test, we generated iSEPs that expressed luciferase and subcutaneously transplantated them into mice. Fifty-six days later, the iSEPs had remained in place, revealing low motility (Extended Data Fig. 8o–q). In summary, in vivo iSEPs formed large epithelial cysts without signs of malignant transformation. They instead behaved like non-malignant epithelial cells with a high proliferative potential.

To molecularly characterize DGTM-reprogrammed cells, RNA sequencing (RNA-seq) analysis was performed on: (1) in vivo iSEPs, (2) in vitro iSEPs (iSEPs generated from $Pdgfra^+$ ADSCs with DGTM-AAV in vitro), (3) $Pdgfra^+$ ADSCs (that is, mesenchymal cells), and (4) primary keratinocytes. When using AAV technologies in dividing cells, the typical goal is to transiently express a transgene in the targeted cells, but AAVs often integrate into the genome[19]. As our DGTM-AAVs express human versions of the DGTM factors, we could compare levels of AAV-derived gene expression (exogenous/human) to the endogenous/mouse genes. In vitro and in vivo iSEPs had high levels of human *DNP63A* expression and variable levels of human-GTM factors (Extended Data Fig. 9a). Levels of endogenous *Trp63* expression were almost completely suppressed in iSEPs, whereas levels of endogenous GTM factors were similar to those of primary keratinocytes. Consistent expression of exogenous factors in iSEPs at different time points suggested the genomic integration of AAV-derived genes.

Clustering analysis of these RNA-seq data indicated that in vivo iSEPs, in vitro iSEPs, and keratinocytes were similar to one another (for example, they expressed high levels of keratinocyte marker genes), but very distinct from Pdgfra$^+$ ADSCs. Transcriptomic analysis revealed that in vivo iSEPs were more similar to primary keratinocytes than in vitro iSEPs. The top gene ontology terms associated with both primary keratinocytes and in vivo iSEPs involved developmental and/or differentiation processes (Extended Data Fig. 9b–h). We compared a time course of gene expression changes for in vivo iSEPs to Pdgfra$^+$ ADSCs and identified five clusters of genes (Extended Data Fig. 9i–l). The time course of gene expression changes may be influenced by the selective survival of clones with stem-like properties, diminishing effects of the wounding itself and resulting inflammation, and epithelial–mesenchymal interactions[8,10,13].

In summary, reprogramming of wound-resident mesenchymal cells enables all regions of the wound to re-epithelialize, relieves the spatial constraints observed during normal healing (Extended Data Fig. 10) and leads to a regenerative functional response in the endogenous skin. Before clinical applications of this potentially transformative, non-surgical technology can be realized, further improvements must be made. These include improving reprogramming efficiency for more prompt epithelialization, and optimizing gene delivery methods to target specific human cell types[20]. Long-term safety studies must also be performed to evaluate off-target effects and tumorigenic potential, especially becuase *DNP63A* is a potential oncogene[21]. The system must also be tested using more clinically relevant animal models of injury. For example, preexisting immunity to AAV reduces transduction efficiencies in humans[22]. Thus, the inflammatory status of the wound, or other clinical features of the patient may affect reprogramming. Finally, a mechanistic understanding of adult stem cell dynamics and plasticity during wound repair is needed, including the extent of dedifferentiation and the inherent heterogeneity of the reprogrammed cells. Our observations constitute an initial proof of principle for functional in vivo regeneration of not only specific and individual cell types, but also of a three dimensional functional tissue in mice. This knowledge might not only be useful for enhancing skin repair, but could also serve to guide in vivo regenerative strategies in other human pathological situations in which tissue or organ homeostasis and repair are impaired.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0477-4.

Received: 30 August 2016; Accepted: 3 August 2018;
Published online 5 September 2018.

1. Gonzales, K. A. U. & Fuchs, E. Skin and its regenerative powers: an alliance between stem cells and their niche. *Dev. Cell* **43**, 387–401 (2017).
2. Sun, B. K., Siprashvili, Z. & Khavari, P. A. Advances in skin grafting and treatment of cutaneous wounds. *Science* **346**, 941–945 (2014).
3. Gurtner, G. C., Werner, S., Barrandon, Y. & Longaker, M. T. Wound repair and regeneration. *Nature* **453**, 314–321 (2008).
4. Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126**, 663–676 (2006).
5. Srivastava, D. & DeWitt, N. In vivo cellular reprogramming: the next generation. *Cell* **166**, 1386–1396 (2016).
6. Driskell, R. R. et al. Distinct fibroblast lineages determine dermal architecture in skin development and repair. *Nature* **504**, 277–281 (2013).
7. Moll, R., Divo, M. & Langbein, L. The human keratins: biology and pathology. *Histochem. Cell Biol.* **129**, 705–733 (2008).
8. Boukamp, P., Breitkreutz, D., Stark, H. J. & Fusenig, N. E. Mesenchyme-mediated and endogenous regulation of growth and differentiation of human skin keratinocytes derived from different body sites. *Differentiation* **44**, 150–161 (1990).
9. Van Muijen, G. N., Warnaar, S. O. & Ponec, M. Differentiation-related changes of cytokeratin expression in cultured keratinocytes and in fetal, newborn, and adult epidermis. *Exp. Cell Res.* **171**, 331–345 (1987).
10. Okazaki, M., Yoshimura, K., Fujiwara, H., Suzuki, Y. & Harii, K. Induction of hard keratin expression in non-nail-matrical keratinocytes by nail-matrical fibroblasts through epithelial–mesenchymal interactions. *Plast. Reconstr. Surg.* **111**, 286–290 (2003).
11. Grimm, D. J. et al. In vitro and in vivo gene therapy vector evolution via multispecies interbreeding and retargeting of adeno-associated viruses. *J. Virol.* **82**, 5887–5911 (2008).
12. Snippert, H. J. et al. Intestinal crypt homeostasis results from neutral competition between symmetrically dividing Lgr5 stem cells. *Cell* **143**, 134–144 (2010).
13. Hirsch, T. et al. Regeneration of the entire human epidermis using transgenic stem cells. *Nature* **551**, 327–332 (2017).
14. De Luca, M., Pellegrini, G. & Green, H. Regeneration of squamous epithelia from stem cells of cultured grafts. *Regen. Med.* **1**, 45–57 (2006).
15. Schmitz, A. et al. Assessing the in vivo epidermal barrier in mice: dye penetration assays. *J. Invest. Dermatol.* **135**, 1–4 (2015).
16. Robson, M. C. et al. The safety and effect of topically applied recombinant basic fibroblast growth factor on the healing of chronic pressure sores. *Ann. Surg.* **216**, 401–408 (1992).
17. Akita, S., Akino, K., Imaizumi, T. & Hirano, A. Basic fibroblast growth factor accelerates and improves second-degree burn wound healing. *Wound Repair Regen.* **16**, 635–641 (2008).
18. Chapman, S., Liu, X., Meyers, C., Schlegel, R. & McBride, A. A. Human keratinocytes are efficiently immortalized by a Rho kinase inhibitor. *J. Clin. Invest.* **120**, 2619–2626 (2010).
19. Chandler, R. J., Sands, M. S. & Venditti, C. P. Recombinant adeno-associated viral integration and genotoxicity: insights from animal models. *Hum. Gene Ther.* **28**, 314–322 (2017).
20. Kotterman, M. A. & Schaffer, D. V. Engineering adeno-associated viruses for clinical gene therapy. *Nat. Rev. Genet.* **15**, 445–451 (2014).
21. Melino, G., Memmi, E. M., Pelicci, P. G. & Bernassola, F. Maintaining epithelial stemness with p63. *Sci. Signal.* **8**, re9 (2015).
22. Louis Jeune, V., Joergensen, J. A., Hajjar, R. J. & Weber, T. Pre-existing anti-adeno-associated virus antibodies as a challenge in AAV gene therapy. *Hum. Gene Ther. Methods* **24**, 59–67 (2013).

Acknowledgements This work was supported by MEXT KAKENHI Grant numbers JP26293381(Grant-in-Aid for Scientific Research (B) to M.K.), JP23689073 (Grant-in-Aid for Young Scientists (A) to M.K.), JP21689046 (Grant-in-Aid for Young Scientists (A) to M.K.), Kyorin University research promotion award to M.K. (2013), JSPS Overseas Research Fellowships (2015–17) to M.K., and the Uehara Memorial Foundation Research Fellowship for Research Abroad (2017–18) to M.K. M.K. thanks H. Green for support materials. T.H. thanks F. Sugiyama for support materials. M.N.S. is supported by NIH-NCI CCSG: P30 014195 and The Leona M. and Harry B. Helmsley Charitable Trust. Work in the laboratory of J.C.I.B. was supported by the G. Harold and Leila Y. Mathers Charitable Foundation, The Leona M. and Harry B. Helmsley Charitable Trust, The Moxie Foundation, The Evergreen Foundation, Fundacion Dr. Pedro Guillen and Universidad Católica San Antonio de Murcia (UCAM).

Reviewer information *Nature* thanks S. Akita, V. Horsley, A. Lombardo and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions M.K., K.H. and J.C.I.B. conceived and designed the experiments. M.K. conceptualized the study and designed and performed most of the experiments and analysis. T.A., T.H., M.S., M.Y. and R.H.-B. prepared the animals and embryos. T.A. and M.Y. helped with histological imaging. K.S., Y.T. and A.S. helped with preparation of plasmid constructs and AAVs. Y.T. and H.E. helped with 3D organotypic cultures. T.H. helped with tumorigenic assay and luciferase detection. M.N.S. analysed the RNA-seq datasets. M.K., D.D.O., J.W., A.O., P.R. and J.C.I.B. prepared the figures and wrote the manuscript. P.M., E.N.D., K.H. and J.C.I.B. coordinated and oversaw the study.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0477-4.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0477-4.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to J.C.I.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Human iSEPs generation. Dermal fibroblasts or adipose-derived stromal cells were seeded at 3,000–5,000 cells/cm^2 in 6-, 12- and 24-well cell culture plates. The next day, retroviral supernatant of DGBM factors was obtained and mixed with complete DMEM medium in a ratio of 4:1 v/v. Polybrene (Sigma) was added at a final concentration of 8 μg/ml. The plate was centrifuged twice at 800*g* for 30 min at 33 °C, then washed with PBS and changed to fresh medium. The medium was changed on days 1, 2 and 4. Cells were reseeded onto mitomycin C-treated 3T3-J2 feeder cells with F medium containing Y27632 as well as vitamin C at a final concentration of 50 μg/ml (Fisher). The medium was changed daily. Typically, epithelial shaped colonies could be identified after 7 to 9 days. When the colonies were ~1,000 cells[23], the cells were passaged by removing the feeder cells and non-converted cells with an initial 1 to 2 min of trypsinization and PBS wash, and a further 5 to 10 min of trypsinization to dissociate the cells within the clones. Passaged cells were maintained in the same way as primary keratinocytes.

Mouse in vitro iSEPs generation. Adipose-derived stromal cells were seeded at 20,000 cells per well in 24-well culture plates. The next day, AAVs were mixed with complete DMEM medium. The medium was changed on days 1, 2 and 4. The medium was changed to F medium containing Y27632 and vitamin C from day 4 or 5. The medium was changed daily. After the emergence of colonies, cells were treated with the same protocols as human iSEPs.

Organotypic culture. Organotypic cultures were prepared using a previously reported method with some modifications[10]. Epithelial cells were cultured in 3D, at the air–liquid interface, on top of a dermal equivalent. Dermal equivalents were constructed by casting 1.0 × 10^6 BJ cells (ATCC CRL-2522) per 6 ml pyruvate (-) complete DMEM medium and 3 ml of bovine dermis-derived atelocollagen, I-PC50 (KOKEN) into a 60-mm Petri dish. The solution was allowed to form a gel and contract for 7 days. Final concentrations of collagen and fibroblasts were 1.7 mg/ml and 1.2 × 10^5 cells/ml, respectively. After loading the dermal equivalents onto an aluminium mount using nylon mesh, an acrylic ring (10 mm in diameter) was set on top and the cells were seeded at more than 4 × 10^5 cells/cm^2 inside the ring with Y27632 containing F-medium. Skin equivalent cultures were maintained in a 60-mm Petri dish or in one well of a six-well plate. For the first 6 days, the concentration of Ca^{2+} inside and outside of the ring was gradually increased to 1.8 mM. On day 7, the volume of the medium inside the ring was reduced to the level of the epithelial cell sheet, so that the epithelial cells were grown at the air–liquid interface. The specimens were harvested on day 14 for histological and immunohistochemical examinations. The schedule for changing the medium during 3D organotypic culture is summarized in Supplementary Table 9.

Clonogenicity assay. The clonogenic ability of primary keratinocytes and epithelial cells generated using DGTM-AAVs was assessed by measuring the proliferative ability of single cell clones. First, the cells were inoculated in 96-well plates at a density of 1 cell per well after being passed through a 70-μm filter. For primary keratinocytes, 3–4 plates were required to obtain around 10 expandable clones, whereas for iSEPs, 1 plate is sufficient. Seven days later, only the wells containing a visually identifiable single colony were passaged to 12-well plates. Seven days later, the plates were stained with rhodanile blue (Alfa Aesar), photographed using a Gel Doc XR+ Imager (Bio-Rad), and analysed for the percentage areas of colonies for each well using ImageJ and Photoshop (Adobe).

Soft-agar colony formation assay. Soft-agar colony formation assay was performed as described elsewhere[24,25]. In brief, 2 ml of 0.5% base layer SeaPlaque agarose (Lonza) in DMEM containing 10% FBS, Keratinocyte feeder medium (KFM) or 3T3-J2 cell-conditioned keratinocyte feeder medium (CKFM) was plated in each well of six-well plates, and 5,000 cells suspended in 1.5 ml 0.5% SeaPlaque agarose (Lonza) in the same medium were placed over the base layer. Medium was replaced every 3 days. Three weeks later the clones were counted and photographed.

Tumorigenicity and teratogenicity assay. In our protocol approved by IACUC, a maximum tumour diameter of <20 mm was permitted and none of the experiments exceeded these limits. mES cells established from C57BL/6-TfN(act-EGFP) mice[26] were a gift from F. Sugiyama. HeLa cells were obtained from ATCC (ATCC CCL-2). Primary keratinocytes were cultured from the back skin of *Krt14*^{cre};*LSL*^{tdTomato} mice. In vivo iSEPs were sorted from primary culture of generated epithelium 3 months after the administration of DGTM-AAVs in *Pdgfra*^{creER};*LSL*^{tdTomato} mice. Cells culture medium (5.0 × 10^6 in 100 μm) were mixed with an equal amount of Matrigel (BD Biosciences) and subcutaneously injected into the back of NOD/SCID IL-2R^{null} (NSG) mice. Twenty-eight days later, after hair removal, the size of the nodule was estimated with the equation; tumour volume (mm^3) = (major axis) × (minor axis)^2/2. Each nodule was collected with overlying skin and histologically investigated. For the assessment of motility and metastatic profile of iSEPs, luciferase-expressing cells were prepared by transduction of MSCV Luciferase PGK-hygro retrovirus (a gift from S. Lowe (Addgene plasmid 18782)) and drug selection. Fifty-six days after injection, the systemic distribution of the cells was investigated by the detection of luciferase.

Animals. C57BL/6 mice, NSG mice, *Pdgfra*^{creER} mice[6], *Pdgfra*^{cre} mice[27], *Krt14*^{cre} mice[28], *Fsp1*^{cre} mice[29], *LSL*^{tdTomato} mice[6], *R26R*^{confetti} mice[12], and *Rosa*$^{mT/mG}$ mice[30] were purchased from the Jackson laboratory.

Mice of both genders were used in this study. All animal procedures were approved by the IACUC committee of the Salk Institute for Biological Studies or the Animal Research Committee of Kyorin University.

Lineage tracing animal. *Pdgfra*^{creER};*LSL*^{tdTomato}, *Pdgfra*^{creER};*Rosa*$^{mT/mG}$, *Pdgfra*^{cre};*LSL*^{tdTomato}, *Pdgfra*^{cre};*Rosa*$^{mT/mG}$, and *Fsp1*^{cre};*LSL*^{tdTomato} mice were used for labelling a broad spectrum of mesenchymal cells to screen for the leakage of signal expression to the epidermis. All strains show aberrant epidermal labelling, but we noticed that a portion of *Pdgfra*^{creER};*LSL*^{tdTomato} mice was free from epidermal leakage. In other words, epidermal leakage with *Pdgfra*^{creER};*LSL*^{tdTomato} mice was specific to individual animals and at different anatomical locations. For a more stringent lineage-tracing system and to eliminate animals with epidermal leakage, we used *Pdgfra*^{creER};*LSL*^{tdTomato} mice. This was accomplished via a two-step selection process: (1) histological investigation of ear biopsies and (2) investigation with primary keratinocytes cultures of skin removed to generate the wound. All lineage-tracing animals were prepared by intraperitoneal injection of 200 μg/day tamoxifen (postnatal day (P)28–32), histological investigation of ear biopsies (P33), and investigation of primary keratinocytes cultures from the skin portion obtained during attachment of the chamber (P35).

Attachment of the chamber and rubber ring. To evaluate the de novo generation of epithelial tissues from the bottom of cutaneous ulcers, we aimed to induce epithelial tissues in skin ulcers that were isolated from the surrounding skin by a skin chamber. To avoid possible contamination from existing epithelial tissues, we performed these assessments using *Krt14*^{cre};*LSL*^{tdTomato} mice. In these mice, cells developmentally expressing *Krt14* are labelled. With the surgical removal of skin and panniculus calnosus, tdTomato signals were detected at the site of the subcutaneously located glands, such as the thyroid and mammary glands. Accordingly, to avoid these structures, two anatomical locations (interscapular area and lateral thoracic area) were chosen as optimal sites for chamber attachment.

Chambers were made up by cutting 1.5-ml Eppendorf tubes. A section was smoothened and broadened by warming with a flame. Two layers of holes were made using 22-gauge needles warmed using a flame. After tanning the internal surface by electric rasp, the chambers were autoclaved. Under general anaesthesia with isoflurane, the site of the chamber attachment was shaved and sterilized. Afterwards, a circle of skin and sub-cutaneous tissue of 1-cm diameter was removed beneath the panniculus carnosus, the chamber was sutured down to the deep fascia with 4–6 horizontal mattress sutures using 4-0 Ethilon (Ethicon Inc.). Then, 200 μl of Cell Matrix Type-1A collagen (Nitta Gelatin) was poured into the gap between the external surface of the chamber and the surrounding skin flap to ensure sealing of the chamber. After suturing the chamber to the skin flap with 4 horizontal mattress sutures, AAVs were administrated and the lid was closed. The edge of the skin flap was glued to the chamber when necessary. Large chambers were made by cutting 5-ml syringes (Becton, Dickinson and Company) using a metal knife warmed with a flame. Basic procedures for manufacturing and fixation were the same as used for the small chamber, except collagen sealing was not used. For the lid, a rubber disk from a syringe was used with a hole in the centre. Fixative rubber rings were made using rubber disks from 5- or 10-ml syringes. After making a hole in the central part, the ring was sutured to the skin and deep fascia to ensure fixation.

Maintenance of the wound. The wound in the chamber is prone to be contaminated because of the use of artificial implant and the closed nature of the wound. When chambers were first attached, in principle, the lids were kept closed to avoid contamination. When chambers were to be changed to larger ones, the lids were kept open for several days in advance, to dry up the wound and prevent carryover of contamination. After changing to the larger chamber, the wound was washed with PBS and photographed every two days under inhalation anaesthesia.

Skin island in chamber assay. An ulcer is generated within a large chamber, but a small patch of skin is left intact in the centre of this ulcer. Four animal groups are defined for comparison; (1) intact skin group, (2) epithelialized skin group, (3) cell sheet groups, and (4) generated epithelium group. For the intact skin group, an island of primary skin was created within a large chamber. For the epithelialized skin group, an ulcer was created and allowed to heal via epithelialization from the surrounding skin (a rubber ring was attached to prevent wound contraction). After epithelialization, the rubber ring was removed and an island of this epithelialized skin was created within a large chamber. For the cell sheet group, epithelial cell sheets were prepared by primary culture of keratinocytes from skin biopsies from each animal. Cell sheets were transplanted onto the backs of these mice via two operations: transplantation of the sheet beneath a silicone sheet, and removal of the overlying skin. An island of this cell sheet-generated epithelium was then created within a large chamber. We attempted to test pure grafted areas, but we found that skin purely composed of a cell-sheet transplant could not survive within a chamber for more than 4 weeks after transplantation. Therefore, we allowed the transplants to contract for another 1–2 weeks before subjecting the cell sheets to the skin island assay. For the DGTM-generated epithelium group, an ulcer was generated using *Pdgfra^{creER};LSL^{tdTomato}* mice and treated via DGTM-AAV administration. After removal of the chamber and confirmation of complete epithelialization, an island of DGTM-generated epithelium was created. For this group, we histologically confirmed that the tested skin island was composed of mesenchymal-reprogrammed epithelium by tdTomato fluorescence after completion of the assay. In each group, on experimental day 0, the skin island was created by circumferential removal of the surrounding skin and the large chamber was attached to make the skin island in an ulcer. Epithelialized areas from the skin island were measured by imaging analysis every 2 days until experimental day 14. For the DGTM-generated epithelium group, after completion of the study period, the tested skin island is subjected to histological analysis to test tdTomato expression in the epithelial layer and confirm that tested skin islands are entirely composed of mesenchymal-derived generated epithelium.

Estimation of nucleated cell number in deep fascia. Deep fascia was dissected out of the chamber, just after the attachment of chamber. Tissues beneath fascia were removed under dissection microscope. Dissected deep fascia was stained with DAPI and mounted on slide glass. The numbers of nucleated cells were counted in ten randomly selected visual fields in 3D images obtained using confocal microscopy. The area of deep fascia was measured in images obtained using a slide scanner. Using the areas and number of nucleated cells, total numbers of nucleated cells were calculated.

Estimation of efficiency of in vivo reprogramming. The efficiency of in vivo reprogramming was estimated by dividing the numbers of clones that contributed to each skin-chamber epithelium in *Pdgfra^{cre};R26R^{confetti}* mice by the numbers of estimated nucleated cells present in the deep fascia at the time of virus administration.

Blood test. Blood samples were obtained from animals that underwent DGTM-AAVs administration and multiple wounding procedures (*Pdgfra^{creER};LSL^{tdTomato}* mice) 3 months after the administration of DGTM-AAVs. Control samples were obtained from age (±2 weeks)-, sex- and genotype-matched non-operated animals. Samples were subjected to test haematology and chemistry assessment at the UCSD murine haematology and coagulation core laboratory.

Vascularity assay. Four groups of animals ($n = 5$) underwent chamber attachment and administration of VEGF165a-AAV (1.0×10^{11} gene copies (GC) per animal), DGTM-AAVs (1.0×10^{11} GC for each factor per animal), GFPNLS-AAVs (4.0×10^{11} GC per animal) or PBS to the inside of the chamber. The volume of solutions are adjusted to 100 µl in all groups. Seven days after application, chambers were collected and immunohistologically investigated for microvessel densities[31]. Sections are stained with H&E and for CD31 and subjected to imaging analysis. Density of vessels is quantified by the proportion of areas of vessels in 6–8 sections of the central portion of the ulcer for each animal. Tissues above and below the fascia are analysed.

Sample collection for RNA-seq analysis. For comparison of primary keratinocytes, iSEPs, and mesenchymal cells, three sets of (1) primary keratinocytes, (2) Pdgfra^{+} ADSCs, (3) in vitro iSEPs, and (4) in vivo iSEPs were prepared from three *Pdgfra^{creER};LSL^{tdTomato}* mice. On postnatal day 35, under general anaesthesia, groin adipose tissues and back skin was collected before chamber attachment and DGTM-AAV inoculation. From skin and adipose tissue, primary keratinocytes (1) and ADSCs were isolated and cultured, respectively. Pdgfra^{+} ADSCs (2) were sorted on the basis of tdTomato signal using a FACS Vantage SE DIVA. Using Pdgfra^{+} ADSCs, in vitro iSEPs (3) were generated with DGTM-AAVs. Mice underwent a stepwise operation to obtain generated epithelial tissues. On postnatal days 125–145, mice were euthanized, and generated epithelium tissues were collected. In vivo iSEPs were isolated with primary keratinocytes from surrounding skin. Pdgfra^{+} in vivo iSEPs (4) were sorted from contaminating primary keratinocytes on the basis of tdTomato signal using a FACS. Other in vivo iSEPs (1 month (1M), $n = 3$, 3M, $n = 2$, 6M, $n = 5$) were obtained from different animals with the same procedures described above. Total RNA was purified from each sample and subjected to RNA-seq analysis.

Toluidine blue penetration assay. Tissue samples were collected in a way that included generated epithelium, ulcer, and the surrounding skin. Samples were dehydrated by incubations (1 min each) in 25%, 50%, and 75% methanol in PBS with a further 1 min in 100% methanol. Then rehydrated with the same series of methanol solutions (1 min incubations) and washed in PBS. Soaking the whole sample caused staining from the backside of the sample and affected the results. Therefore, we attached the tube on the surface of the sample to partially include the surrounding skin, and added 0.1% toluidine blue O/PBS into the tube for 1 min before washing[32]. Samples were photographed. The frozen cross-section was also photographed. After sectioning, both non-stained sections and serial H&E sections were prepared to analyse the detailed location of toluidine blue stain.

Lucifer yellow dye penetration assay. Under general anaesthesia, mice were restrained in Petri dishes with generated tissues in contact with 1 mM Lucifer yellow in PBS (pH 7.4) at room temperature. After 1 h of incubation, mice were euthanized and tissues were dissected, frozen and sectioned at a thickness of 5 µm[33]. The sections were counterstained with 10 µg/ml DAPI and then analysed by fluorescence microscopy.

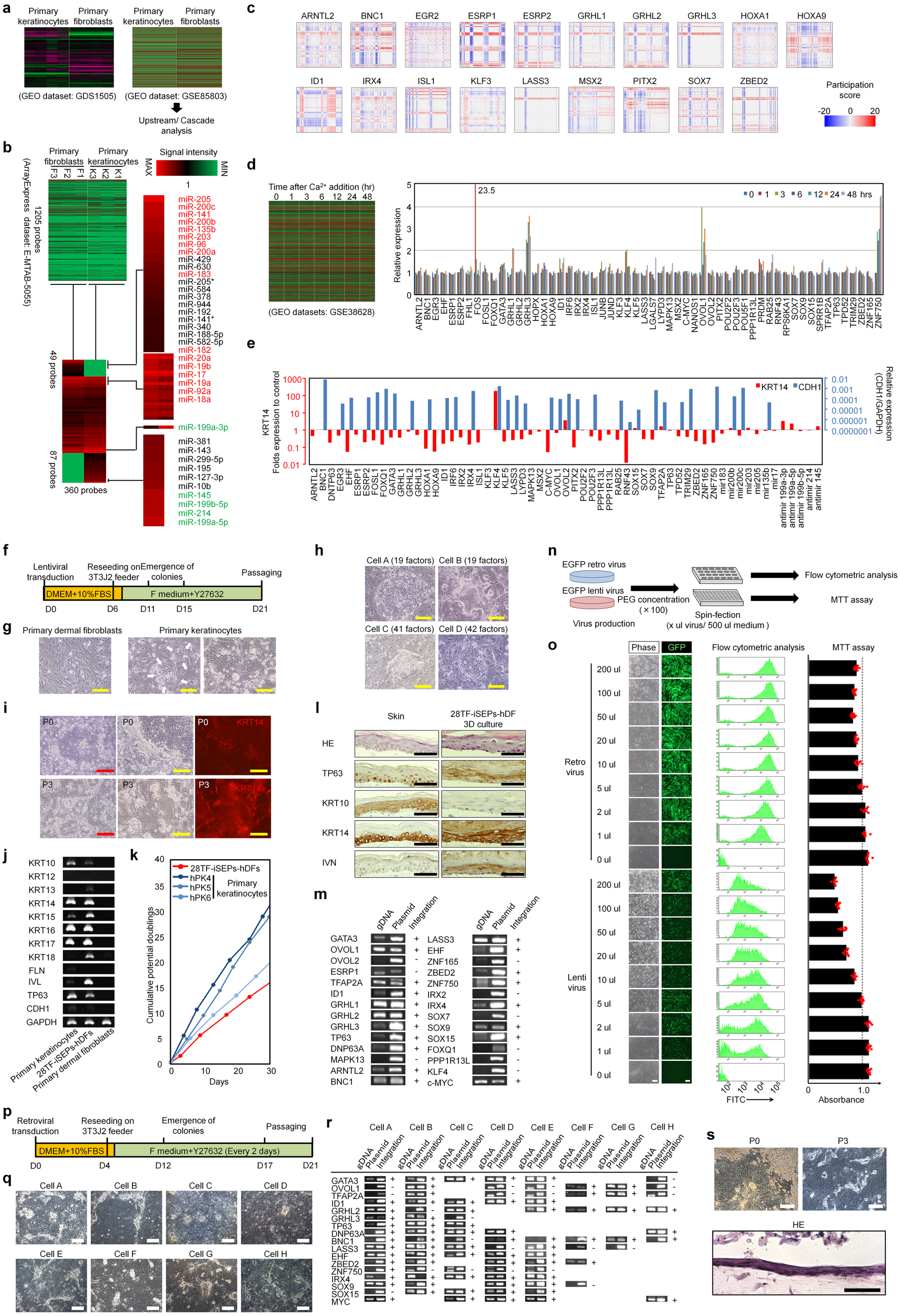
Transepidermal water loss (TEWL) measurement. On the day of measurement, hair was carefully shaved from the testing area in mice under general anaesthesia using isoflurane. Special effort was taken not to damage the skin. TEWL was measured using a VapoMeter (Delfin Technologies) with nail adapters at the area of generated epithelium and intact skin. Because TEWL measurements can be influenced by factors such as room and body temperatures and depth of anaesthesia, at least three rounds of measurements were taken. Subsequently, a portion of the skin was surgically removed to generate an ulcer (1-cm diameter) and the TEWL value was measured more than three times.

Statistical analysis. No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

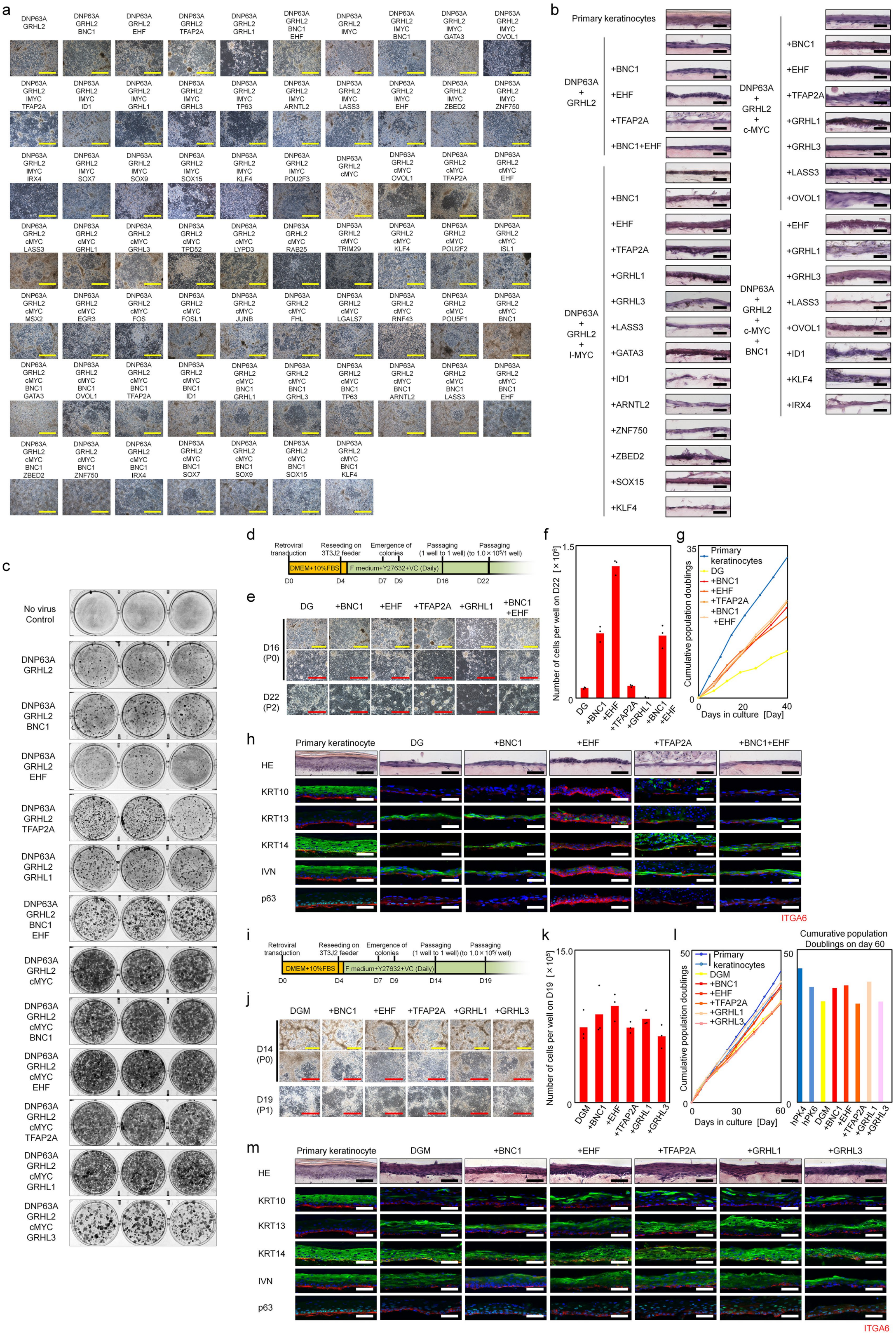
Data availability. Data from the microarray, RNA-seq and miRNA microarray analyses have been deposited in the Gene Expression Omnibus (GEO) and ArrayExpress under accession numbers GSE85803, GSE106419 and E-MTAB-5055, respectively. All other relevant data that support the findings of this study are available from the corresponding author upon reasonable request.

23. Rheinwatd, J. G. & Green, H. Serial cultivation of strains of human epidermal keratinocytes: the formation of keratinizing colonies from single cells. *Cell* **6**, 331–343 (1975).
24. Harajly, M. et al. p53 restoration in induction and maintenance of senescence: differential effects in premalignant and malignant tumor cells. *Mol. Cell. Biol.* **36**, 438–451 (2016).
25. Provost, J. J. et al. Urokinase plasminogen activator receptor induced non-small cell lung cancer invasion and metastasis requires NHE1 transporter expression and transport activity. *Cell Oncol. (Dordr.)* **35**, 95–110 (2012).
26. Shimizukawa, R. et al. Establishment of a new embryonic stem cell line derived from C57BL/6 mouse expressing EGFP ubiquitously. *Genesis* **42**, 47–52 (2005).
27. Roesch, K. et al. The transcriptome of retinal Müller glial cells. *J. Comp. Neurol.* **509**, 225–238 (2008).
28. Dassule, H. R., Lewis, P., Bei, M., Maas, R. & McMahon, A. P. Sonic hedgehog regulates growth and morphogenesis of the tooth. *Development* **127**, 4775–4785 (2000).
29. Song, K. et al. Heart repair by reprogramming non-myocytes with cardiac transcription factors. *Nature* **485**, 599–604 (2012).
30. Muzumdar, M. D., Tasic, B., Miyamichi, K., Li, L. & Luo, L. A global double-fluorescent Cre reporter mouse. *Genesis* **45**, 593–605 (2007).
31. Kurashige, C. et al. Roles of receptor activity-modifying protein 1 in angiogenesis and lymphangiogenesis during skin wound healing in mice. *FASEB J.* **28**, 1237–1247 (2014).
32. Hardman, M. J., Sisi, P., Banbury, D. N. & Byrne, C. Patterned acquisition of skin barrier function during development. *Development* **125**, 1541–1552 (1998).
33. Matsuki, M. et al. Defective stratum corneum and early neonatal death in mice lacking the gene for transglutaminase 1 (keratinocyte transglutaminase). *Proc. Natl Acad. Sci. USA* **95**, 1044–1049 (1998).



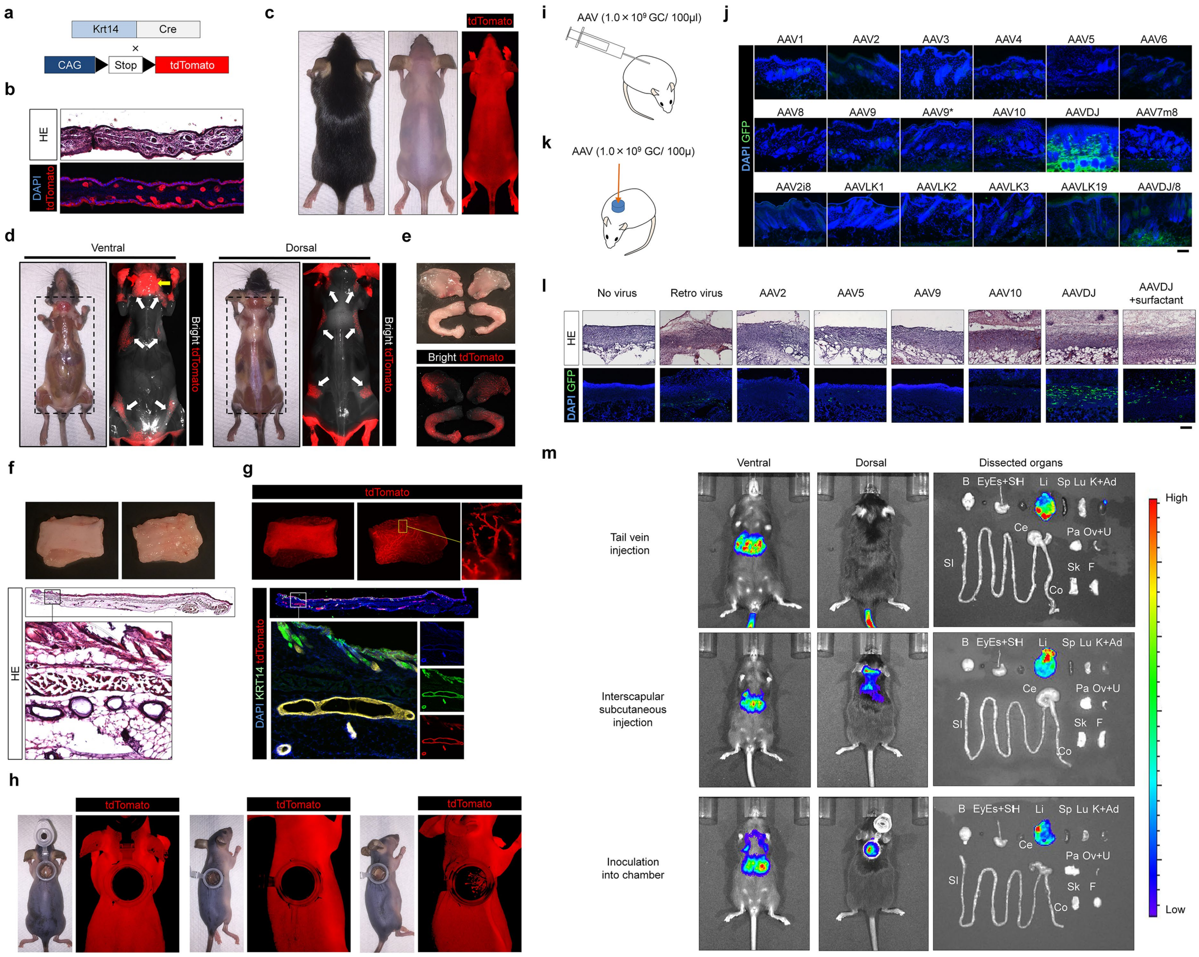
Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Selection of factors for reprogramming of human mesenchymal cells into iSEPs. a, Gene-expression analyses from DNA microarrays of human primary keratinocytes and primary hDFs. Pink or red indicates higher expression levels, green indicates lower expression levels. **b**, MicroRNA microarray analyses from three pairs of primary human keratinocytes and primary hDFs obtained from different tissues of a single subject. **c**, Reversal participation analysis of a candidate gene set (19 genes) for keratinocyte specification. **d**, Changes in the expression levels of candidate transcription factors after calcium-induced terminal differentiation of keratinocytes. **e**, Changes in the expression levels of keratinocyte markers after transduction of each candidate gene as assessed by quantitative PCR (qPCR). **f**, Schematic of the experimental design for the generation of iSEPs from primary hDFs. **g**, Morphological analysis of primary hDFs and human primary keratinocytes. Arrows indicate keratinocyte colonies on feeder cells. Scale bars, 200 µm. Similar results were observed five times for both cell types. **h**, Representative bright field images showing the morphology of colonies obtained during the selection of factors for the generation of iSEPs. Combinations of factors were optimized on the basis of colony morphology. Scale bars, 200 µm. **i**, Representative bright field images (left and middle) showing the morphology of colonies at passage 0 (P0) and passage 3 (P3) obtained after the transduction of 28 factors. The KRT14-RFP vector was transduced with the 28 factors (right). Red scale bars, 500 µm; yellow scale bars, 200 µm. **j**, RT–PCR analyses of keratinocyte markers. **h**, **i**, **j**, Images are from one experiment. **k**, Growth curve of primary keratinocytes and 28TF-iSEPs on feeder cells. Data are mean of technical triplicates. **l**, Representative images showing H&E staining of human skin and 3D organotypic culture of 28TF-iSEPs-hDF. Scale bars, 100 µm. Similar results were observed in two organotypic cultures. **m**, Transgene-specific PCR analyses of genomic DNA of 28TF-iSEPs. Each plasmid was used for comparative controls. Images are from one experiment. **n**, Schematic of the experiment for comparative assessment of transduction efficiency and cytotoxicity between concentrated enhanced GFP (eGFP)-expressing retroviruses and lentiviruses. **o**, Higher eGFP expression could be obtained with lower cytotoxicity with retroviruses than lentiviruses. Consistent findings were observed in three technical replicates for microscopic findings and flow cytometric analyses. Results of MTT (3-[4,5-dimethylthiazol-2-yl]-2,5 diphenyl tetrazolium bromide) assay represent the mean of three technical replicates. Overlaid dot plots indicate the distribution of the data. **p**, Schematic of the experiment for generation of iSEPs-ADSCs. **q**, Representative bright field images showing colony morphologies during factor reduction. Scale bars, 200 µm. **r**, Transgene-specific PCR analyses of genomic DNA. Plasmids were used as controls. Integration of plasmid-derived sequences is described. **q**, **r**, Images are from one experiment. **s**, Representative images showing the morphological analysis and H&E staining of 3D organotypic culture of iSEPs-ADSCs generated by the transduction of *DNP63A* and *GRHL2*. White scale bars, 200 µm; black scale bar, 100 µm. Similar results were confirmed in two independent experiments. **j**, **m**, **r**, For gel source data, see Supplementary Fig. 1. **o**, For gating strategy example, see Supplementary Fig. 2.

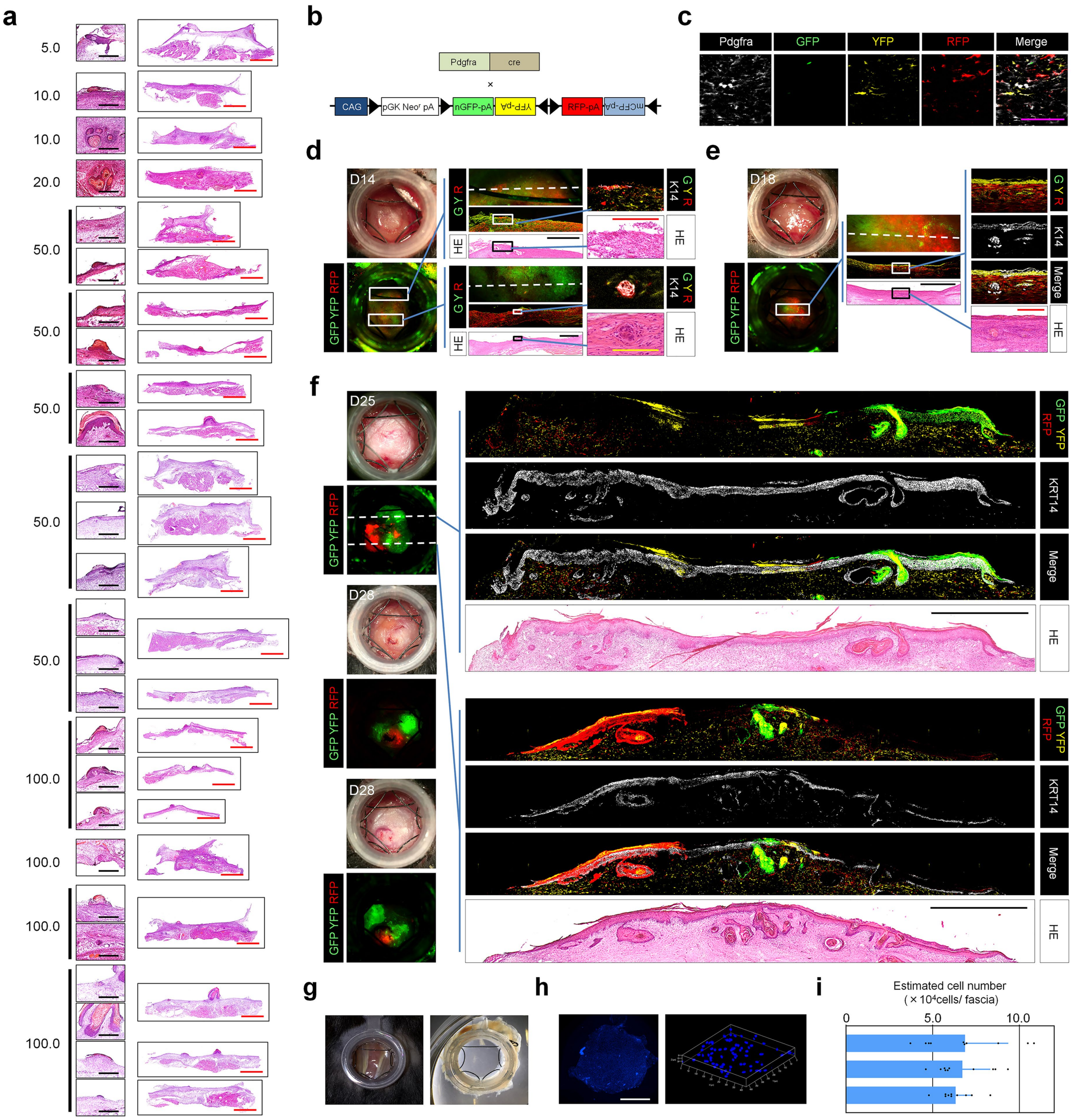

a
b
Primary keratinocytes
DNP63A + GRHL2
+BNC1
+EHF
+TFAP2A
+BNC1+EHF
DNP63A + GRHL2 + l-MYC
+GRHL1
+GRHL3
+LASS3
+GATA3
+ID1
+ARNTL2
+ZNF750
+ZBED2
+SOX15
+KLF4
DNP63A + GRHL2 + c-MYC
+OVOL1
DNP63A + GRHL2 + c-MYC + BNC1
+IRX4
c
No virus Control
d
Retroviral transduction
Reseeding on 3T3J2 feeder
Emergence of colonies
Passaging (1 well to 1 well)
Passaging (to 1.0 × 10⁵/1 well)
DMEM+10%FBS
F medium+Y27632+VC (Daily)
D0
D4
D7
D9
D16
D22
e
DG
D16 (P0)
D22 (P2)
f
Number of cells per well on D22 [× 10⁶]
g
Cumulative population doublings
Days in culture [Day]
h
HE
KRT10
KRT13
KRT14
IVN
p63
ITGA6
i
DMEM+10%FBS
F medium+Y27632+VC (Daily)
D14
D19
j
DGM
D14 (P0)
D19 (P1)
k
Number of cells per well on D19 [× 10⁶]
l
Cumurative population Doublings on day 60
hPK4
hPK6
m

Extended Data Fig. 2 | See next page for caption.

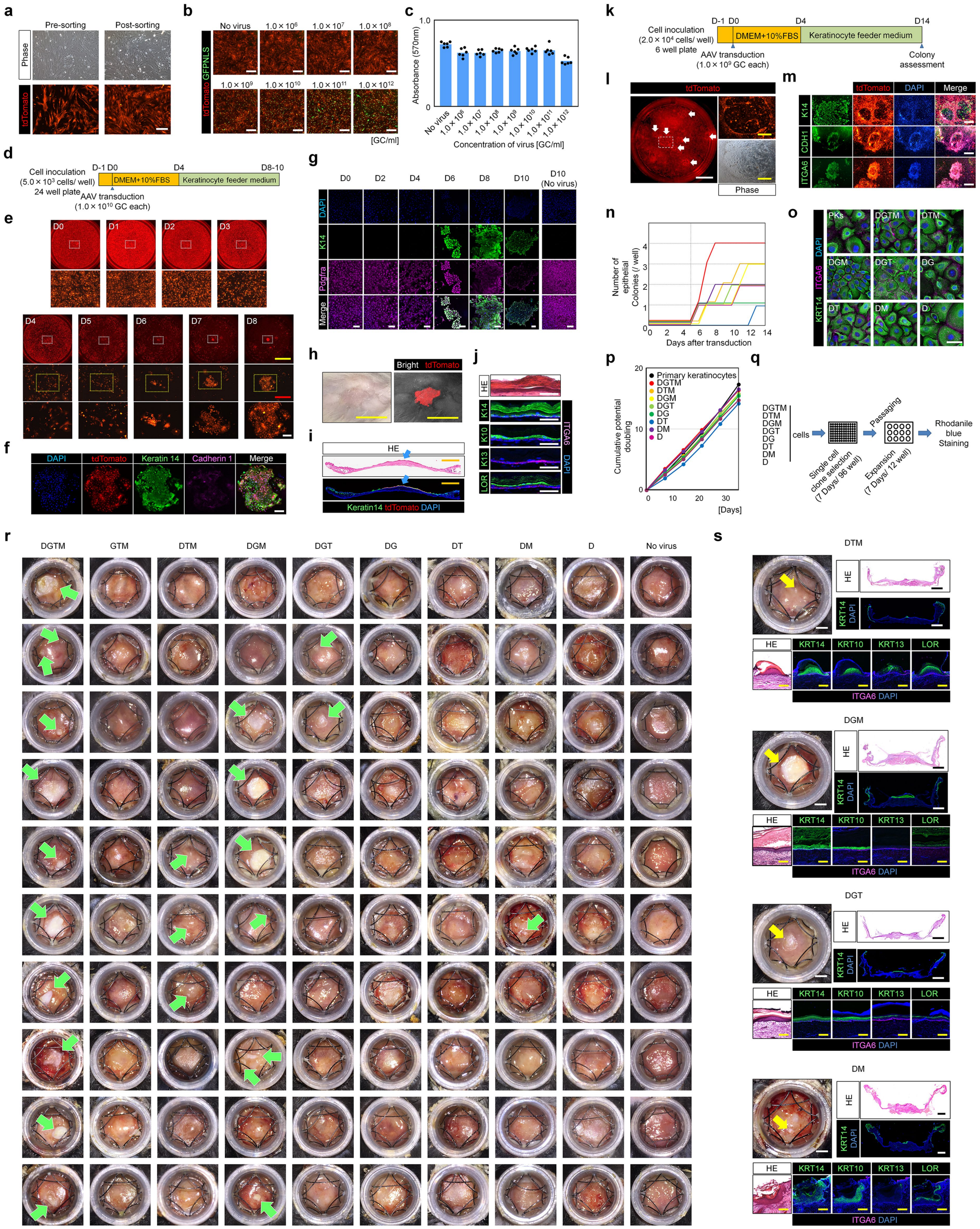
Extended Data Fig. 2 | Optimization of factors for reprogramming of hADSCs into iSEPs. **a**, Bright field images showing the representative morphology of colonies that emerged during generation of iSEPs after transduction of 67 combinations of factors tested in addition to the minimum factors *DNP63A* and *GRHL2*. Scale bars, 1,000 µm. **b**, Representative H&E staining images of 3D organotypic cultures of iSEPs induced with 35 combinations of factors. Scale bars, 100 µm. **c**, Rhodanile staining of cells generated with selected combinations of factors 14 days after transduction. **d**, Schematic representation of the experiment for testing the combinations of DG and non-MYC factors. **e**, Representative bright field images showing cell morphologies of generated iSEPs. Yellow scale bars, 1,000 µm; red scale bars, 500 µm. **f**, Quantification of iSEPs generated with different combinations of factors on day 22. Values represent the mean of three technical replicates. Overlaid dot plots indicate the distribution of the data. **g**, Growth curves of iSEPs and human primary keratinocytes. Cells with GRHL1 could not be isolated. Data are average of triplicates. **h**, H&E staining and immunohistochemical analysis of 3D organotypic cultures. Scale bars, 100 µm. **i**, Schematic representation of the experiment for testing the combinations of DGM+1 factors. **j**, Representative bright field images showing cell morphologies of generated iSEPs. Yellow scale bars, 1,000 µm; red scale bars, 500 µm. **k**, Quantification of iSEPs generated with different combinations of factors on day 19. Values represent mean of three technical replicates. Overlaid dot plots indicate the distribution of the data. **l**, Growth curves of iSEPs and human primary keratinocytes (hPK1 and hPK4), and the cumulative population doublings of iSEPs on day 60. Data are the mean of triplicates. **m**, H&E staining and immunohistochemical analysis of 3D organotypic cultures. Scale bars, 100 µm. **a**, **b**, **h**, **m**, Findings were confirmed in two technical replicates. **c**, **e**, **j**, Findings were confirmed in two independent experiments.


a
Krt14
Cre
×
CAG
Stop
tdTomato
b
HE
DAPI tdTomato
c
tdTomato
d
Ventral
Dorsal
Bright tdTomato
e
Bright tdTomato
f
HE
g
tdTomato
DAPI KRT14 tdTomato
h
tdTomato
i
AAV (1.0 × 10⁹ GC/ 100μl)
k
AAV (1.0 × 10⁹ GC/ 100μ)
j
DAPI GFP
AAV1
AAV2
AAV3
AAV4
AAV5
AAV6
AAV8
AAV9
AAV9*
AAV10
AAVDJ
AAV7m8
AAV2i8
AAVLK1
AAVLK2
AAVLK3
AAVLK19
AAVDJ/8
l
No virus
Retro virus
AAV2
AAV5
AAV9
AAV10
AAVDJ
AAVDJ +surfactant
HE
DAPI GFP
m
Ventral
Dorsal
Dissected organs
Tail vein injection
Interscapular subcutaneous injection
Inoculation into chamber
B
EyEs+StH
Li
Sp
Lu
K+Ad
Ce
Pa
Ov+U
SI
Sk
F
Co
High
Low

Extended Data Fig. 3 | Defining the procedure of in vivo experiments. **a**, Lineage tracing using $\mathit{Krt14}^{cre};\mathit{LSL}^{tdTomato}$ mice. **b**, H&E staining and immunohistochemical analysis of ear biopsies from $\mathit{Krt14}^{cre};\mathit{LSL}^{tdTomato}$ mice. **c**, Exterior and stereoscopic view of $\mathit{Krt14}^{cre};\mathit{LSL}^{tdTomato}$ mice. **d**, Appearance and stereoscopic view of $\mathit{Krt14}^{cre};\mathit{LSL}^{tdTomato}$ mice after resection of skin. White arrows indicate signals from glands such as the thyroid and mammary glands. **e**, Appearance and stereoscopic view of resected mammary gland. **f**, Appearance and H&E staining of resected axillary skin and subcutaneous tissues including the mammary gland. **g**, Stereoscopic and immunohistochemical analysis of resected skin and subcutaneous tissues including the mammary gland. **h**, Appearance and stereoscopic analysis after chamber attachment. tdTomato signals were not detected inside the chamber when attached at the interscapular area and lateral thoracic area, whereas signals were detected when attached close to the axilla. The procedure was optimized not to include the mammary gland within the wound chamber. **i**, Schematic representation of the subcutaneous injection of AAVs into the back skin of mice. Eighteen serotypes of AAVs expressing eGFP under the CAG promoter were tested. **j**, Fluorescence images showing GFP signals of the injected skin and subcutaneous tissue 1 week after the injection of eGFP-expressing AAVs of different serotypes. AAVDJ resulted in the highest levels of GFP fluorescence. Scale bar, 200 μm. **k**, Schematic representation of the inoculation of AAVs in a skin chamber on the back of a mouse. Tested AAVs included: (1) AAVDJ with or without the addition of surfactant, (2) AAVs previously identified as optimal for skin and wounds (AAV2 and AAV5), and (3) new AAV serotypes (AAV9 and AAV10). Highly concentrated retroviruses were used as a control. **l**, H&E and fluorescence analyses of ulcers 1 week after the inoculation of eGFP-expressing viruses. AAVDJ without surfactant yielded by far the highest gene transduction efficiency. Scale bar, 200 μm. **m**, In vivo luminescent images 1 week after the administration of luciferase-expressing AAVs through tail vein injection, interscapular subcutaneous injection or inoculation into a chamber on the back. Systemic directivity of AAVDJ is mainly focused to the liver. B, brain; Ce, caecum; Co, colon; Es + St, oesophagus and stomach; Ey, eye; F, fat; H, heart; K + Ad, kidney and adrenal gland; L, liver; Lu, lung; Ov + U, ovary and uterus; Pa, pancreas; SI, small intestine; Sk, skin; Sp, spleen. Similar findings were obtained in five (**b**), three (**c**–**g**, **l**) and two (**h**, **j**, **m**) independent experiments.

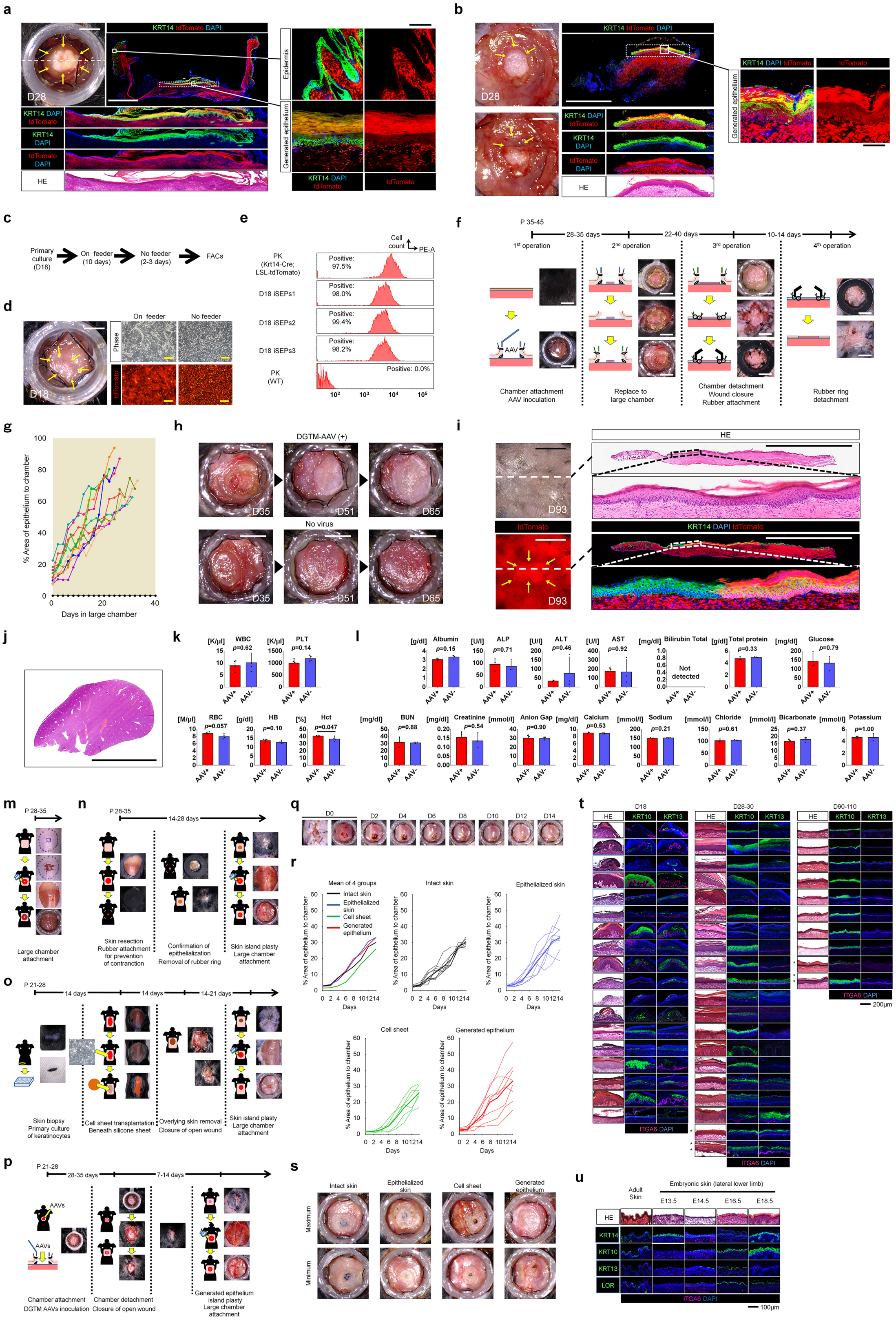

a
5.0
10.0
10.0
20.0
50.0
50.0
50.0
50.0
50.0
100.0
100.0
100.0
100.0
b
Pdgfra
cre
×
CAG
pGK Neo^r pA
nGFP-pA
YFP-pA
RFP-pA
mCFP-pA
c
Pdgfra
GFP
YFP
RFP
Merge
d
D14
GFP YFP RFP
G Y R
HE
G Y R K14
e
D18
GFP YFP RFP
G Y R
K14
Merge
HE
f
D25
D28
D28
GFP YFP RFP
KRT14
Merge
HE
g
h
i
Estimated cell number
(× 10⁴cells/ fascia)
0
5.0
10.0

Extended Data Fig. 4 | Generation of epithelial tissue after DGTM-AAV administration in vivo. a, H&E analysis of epithelial tissue 18 days after DGTM-AAV administration. Both high magnification (left) and low magnification images (right) are shown. Viral titres are indicated on the left (×10^{10} GC/animal). Black scale bars, 500 μm; red scale bars, 2 mm. Data from 14 out of 25 animals treated with different titres of DGTM-AAV as shown (summarized data is shown in Fig. 1d). **b**, Lineage tracing using *Pdgfra*^{cre};*R26R*^{confetti} mice. **c**, A representative image showing GFP, YFP, RFP fluorescence and Pdgfra staining of the subcutaneous tissue isolated from the back skin of a *Pdgfra*^{cre};*R26R*^{confetti} mouse. Mesenchymal cells were differentially labelled with GFP, YFP, RFP or not labelled. Similar findings were confirmed in five animals. **c**, Representative visual, stereoscopic and histological images showing single-cell-derived epithelial cell clusters 14 days after DGTM-AAVs administration. Similar findings were confirmed in two animals (out of three animals treated with the same procedure). **d**, **e**, Representative visual, stereoscopic, and histological images showing participation of YFP-labelled and non-labelled epithelial cell clusters to a single epithelial tissue on day 14 (**d**) and day 18 (**e**). Similar findings were confirmed in all three animals treated with the same procedure for each day. **f**, Representative visual and stereoscopic images of three mice (left) showing that multicolour-labelled cell clusters contribute to epithelial tissues covering the ulcer surface inside the chamber. Histological images (right) of one animal showing the participation of GFP, YFP, RFP and unlabelled cells to generated epithelial tissue. Similar findings were confirmed in all three animals. **g**, For the estimation of cell number in deep fascia at the time of surgery, deep fascia was dissected out just after the attachment of the chamber. **h**, Stereoscopic images showing a horizontal image of the deep fascia after DAPI staining (left). Imaging analysis for the counting of nucleated cells (right). **g**, **h**, All three samples were processed with similar findings. **i**, Estimated cell numbers in the deep fascia just after the attachment of chamber for three animals. Data represents mean ± s.d. of from ten stereoscopic images. Overlaid dot plots indicate the distribution of the data. **c**–**h**, White scale bar, 5 mm; black scale bars, 1 mm; red scale bars, 200 μm; yellow and magenta scale bars, 100 μm. External diameter of the chamber, 12.5 mm.


a
Pre-sorting
Post-sorting
Phase
tdTomato
b
No virus
1.0 × 10^6
1.0 × 10^7
1.0 × 10^8
1.0 × 10^9
1.0 × 10^10
1.0 × 10^11
1.0 × 10^12
tdTomato GFPNLS
[GC/ml]
c
Absorbance (570nm)
1.0
0
No virus
Concentration of virus [GC/ml]
d
Cell inoculation
(5.0 × 10^3 cells/ well)
24 well plate
D-1 D0
D4
D8-10
DMEM+10%FBS
Keratinocyte feeder medium
AAV transduction
(1.0 × 10^10 GC each)
e
D0
D1
D2
D3
D4
D5
D6
D7
D8
f
DAPI
tdTomato
Keratin 14
Cadherin 1
Merge
g
D0
D2
D4
D6
D8
D10
D10 (No virus)
DAPI
K14
Pdgfra
Merge
h
Bright
tdTomato
i
HE
Keratin14 tdTomato DAPI
j
HE
K14
K10
K13
LOR
ITGA6 DAPI
k
Cell inoculation
(2.0 × 10^4 cells/ well)
6 well plate
D-1 D0
D4
D14
DMEM+10%FBS
Keratinocyte feeder medium
AAV transduction
(1.0 × 10^9 GC each)
Colony assessment
l
tdTomato
Phase
m
tdTomato
DAPI
Merge
K14
CDH1
ITGA6
n
Number of epithelial Colonies (/ well)
Days after transduction
0
2
4
6
8
10
12
14
o
PKs
DGTM
DTM
DGM
DGT
DG
DT
DM
D
DAPI
ITGA6
KRT14
p
Cumulative potential doubling
[Days]
Primary keratinocytes
DGTM
DTM
DGM
DGT
DG
DT
DM
D
0
10
20
30
q
DGTM
DTM
DGM
DGT
DG
DT
DM
D
cells
Single cell clone selection (7 Days/ 96 well)
Passaging
Expansion (7 Days/ 12 well)
Rhodanile blue Staining
r
DGTM
GTM
DTM
DGM
DGT
DG
DT
DM
D
No virus
s
DTM
DGM
DGT
DM
HE
KRT14 DAPI
KRT14
KRT10
KRT13
LOR
ITGA6 DAPI

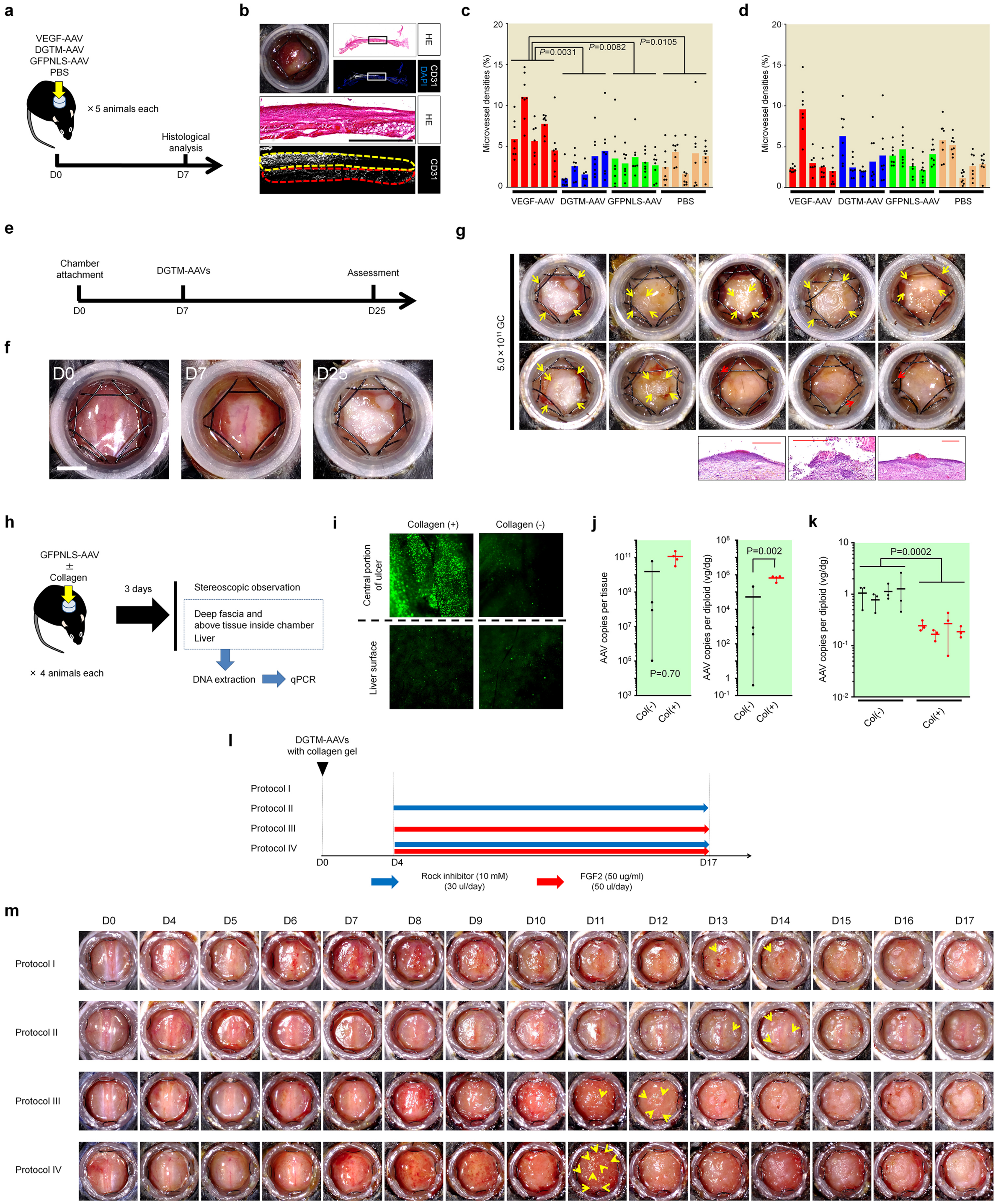
Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | Reprogramming of mouse mesenchymal cells with different combinations of DGTM-AAVs. a, Mesenchymal cells were sorted out from mouse adipose-derived stromal cell fractions by *Pdgfra*^{creER}-driven tdTomato signals with the proportion of 63.2 ± 16.0% ($n = 12$, established from different animals). For gating strategy example, see Supplementary Fig. 2. Scale bar, 200 µm. **b**, Fluorescence images of the cells 3 days after addition of GFPNLS-AAV with indicated titre in vitro. GFPNLS expression increases with increase of titre. Scale bars, 200 µm. Similar findings were confirmed in two experiments. **c**, Results of MTT cell-viability assay of the cells 3 days after the addition of GFPNLS-AAV at the indicated titre in vitro. Overlaid dot plots indicate the distribution of the data ($n = 6$, technical replicates). **d**, Schematic of the experimental design. **e**, Time course stereoscopic and fluoroscopic analysis of the emergence of an iSEPs colony. Yellow scale bar, 5 mm; red scale bar, 1 mm; white scale bar, 200 µm. **f**, Immunocytochemical analysis of the iSEPs colony in **e**. **e**, **f**, Similar findings were confirmed in two sets of eight wells of samples. **g**, Time course immunocytochemical analysis of iSEPs colony emergence. iSEPs colonies are strongly positive for Krt14 and Pdgfra on initial emergence. With time, Pdgfra signal intensity decreases. Scale bars, 200 µm. Similar findings were confirmed in a set of eight wells of samples. **h**, Appearance (left) and stereoscopic analysis (right) of transplanted cell sheet. tdTomato signals indicates the area of cell sheet survival. Yellow scale bars, 1.0 cm. **i**, Findings of H&E and immunohistological analysis of transplanted cell sheet. Arrows indicate the position of magnified findings in **j**. Orange scale bars, 2 mm. **j**, H&E and immunohistological analysis of transplanted cell sheet. White scale bars, 100 µm. **h**, **i**, **j**, Similar findings were confirmed in five animals in two sets of experiments. **k**, Schematic representation of the experimental design at the lower titre of virus. **l**, DGTM-AAVs transduced wells. Arrows indicates epithelial shaped colonies. The dotted outline is shown magnified on the right. White scale bar, 5 mm; yellow scale bars, 1 mm. **m**, Immunocytochemical analysis of epithelial shaped colonies. White scale bars, 1 mm. **l**, **m**, Similar findings were confirmed in two sets of three wells (six-well plate). **n**, Time course numbers of epithelial colonies in eight wells (24-well plate) treated with DGTM-AAVs. Similar findings were confirmed in two series of experiments. **o**, Immunocytochemical analysis of epithelial cells obtained after transduction of different combinations of factors into Pdgfra^{+} mADSCs. Scale bar, 50 µm. Images are representative of one experiment. **p**, Proliferation of epithelial cells obtained after transduction of different combinations of factors into Pdgfra^{+} mADSCs. Data are mean of triplicates. **q**, Schematic representation of clonogenicity assessment by expansion of single-cell-derived clones. **r**, Analysis of ulcers 28 days after administration of nine combinations of AAV and no virus control (ten animals for each group, including samples shown in Fig. 1b, e). Green arrows indicate epithelial tissues confirmed by histological analysis. External diameter of chamber, 12.5 mm. Data are summarized in Fig. 2c–e. Findings were confirmed in a set of experiment. **s**, Epithelial tissue generated with DTM-, DGM-, DGT-, and DM-AAVs in vivo. Arrows indicate generated epithelial tissue. Black and white scale bars, 2 mm; yellow scale bars, 200 µm. Similar findings were confirmed in three (DTM), six (DGM), two (DGT) or one (DM) animals (out of ten animals for each group).


a
D28
KRT14 tdTomato DAPI
Epidermis
Generated epithelium
KRT14 DAPI tdTomato
KRT14 DAPI
tdTomato DAPI
HE
tdTomato
b
D28
KRT14 tdTomato DAPI
KRT14 DAPI tdTomato
tdTomato
Generated epithelium
HE
c
Primary culture (D18)
On feeder (10 days)
No feeder (2-3 days)
FACs
d
D18
On feeder
No feeder
Phase
tdTomato
e
Cell count
PE-A
PK (Krt14-Cre; LSL-tdTomato)
Positive: 97.5%
D18 iSEPs1
Positive: 98.0%
D18 iSEPs2
Positive: 99.4%
D18 iSEPs3
Positive: 98.2%
PK (WT)
Positive: 0.0%
10²
10³
10⁴
10⁵
f
P 35-45
28-35 days
22-40 days
10-14 days
1st operation
2nd operation
3rd operation
4th operation
AAV
Chamber attachment AAV inoculation
Replace to large chamber
Chamber detachment Wound closure Rubber attachment
Rubber ring detachment
g
% Area of epithelium to chamber
Days in large chamber
0
20
40
60
80
100
h
DGTM-AAV (+)
No virus
D35
D51
D65
i
HE
D93
tdTomato
KRT14 DAPI tdTomato
j
k
WBC
PLT
RBC
HB
Hct
p=0.62
p=0.14
p=0.057
p=0.10
p=0.047
AAV+
AAV-
l
Albumin
ALP
ALT
AST
Bilirubin Total
Total protein
Glucose
p=0.15
p=0.71
p=0.46
p=0.92
Not detected
p=0.33
p=0.79
BUN
Creatinine
Anion Gap
Calcium
Sodium
Chloride
Bicarbonate
Potassium
p=0.88
p=0.54
p=0.90
p=0.53
p=0.21
p=0.61
p=0.37
p=1.00
m
P 28-35
Large chamber attachment
n
P 28-35
14-28 days
Skin resection Rubber attachment for prevention of contraction
Confirmation of epithelialization Removal of rubber ring
Skin island plasty Large chamber attachment
o
P 21-28
14 days
14 days
14-21 days
Skin biopsy Primary culture of keratinocytes
Cell sheet transplantation Beneath silicone sheet
Overlying skin removal Closure of open wound
Skin island plasty Large chamber attachment
p
P 21-28
28-35 days
7-14 days
AAVs
Chamber attachment DGTM AAVs inoculation
Chamber detachment Closure of open wound
Generated epithelium island plasty Large chamber attachment
q
D0
D2
D4
D6
D8
D10
D12
D14
r
Mean of 4 groups
Intact skin
Epithelialized skin
Cell sheet
Generated epithelium
% Area of epithelium to chamber
Days
0 2 4 6 8 10 12 14
s
Intact skin
Epithelialized skin
Cell sheet
Generated epithelium
Maximum
Minimum
t
D18
D28-30
D90-110
HE
KRT10
KRT13
ITGA6 DAPI
200μm
u
Adult Skin
Embryonic skin (lateral lower limb)
E13.5
E14.5
E16.5
E18.5
HE
KRT14
KRT10
KRT13
LOR
ITGA6 DAPI
100μm

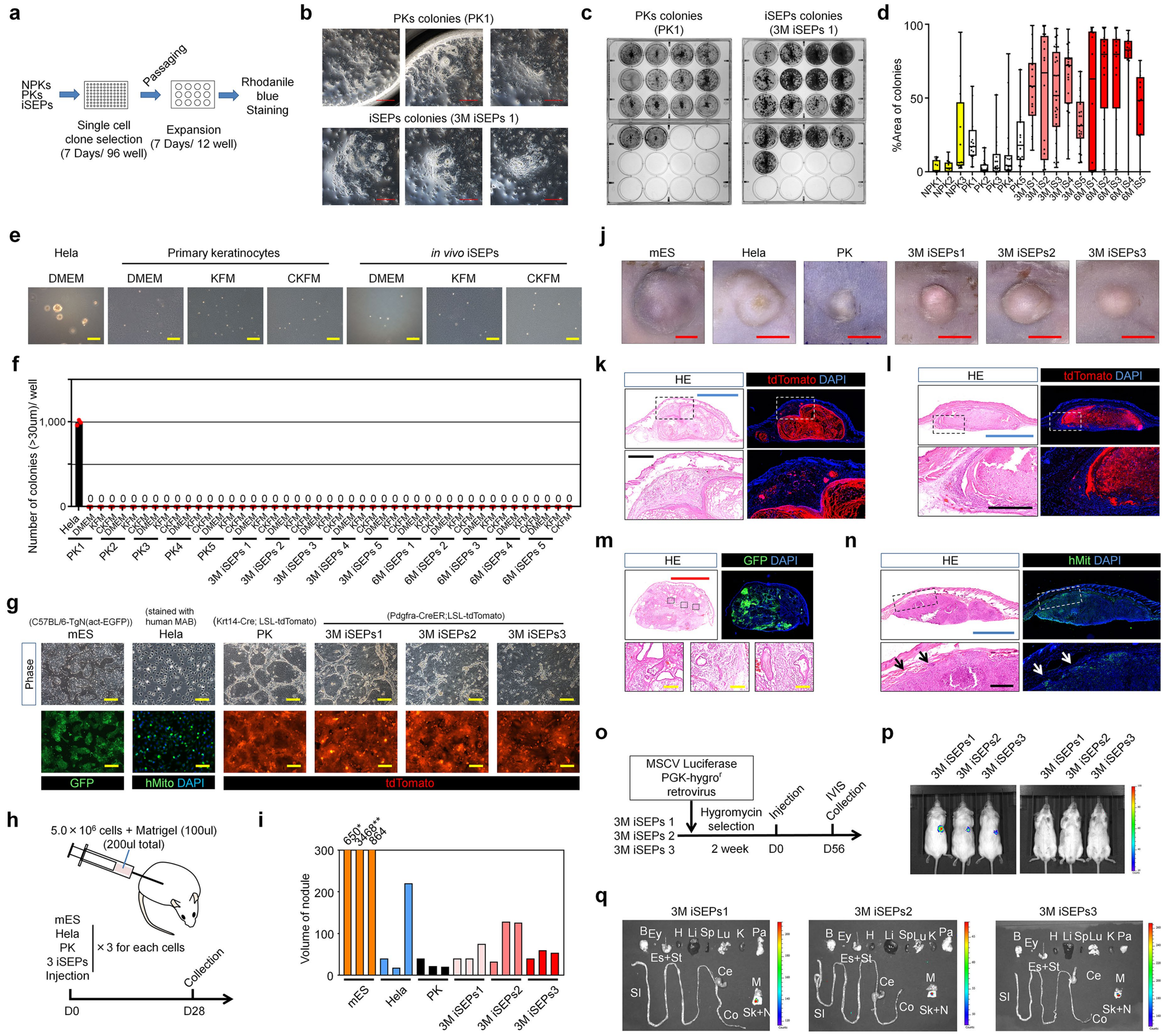
Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | DGTM-AAVs enable generation of epithelium with the ability to cover ulcers. a, Appearance (top left) and immunohistochemical analysis (top middle) of ulcers, 28 days after the administration of DGTM-AAV in *Pdgfra*^{creER};*LSL*^{tdTomato} mice, analysed for KRT14 and tdTomato expression. Yellow arrows indicate the periphery of the generated epithelium. The white dotted line indicates the approximate position of the histological section. The white dotted outline indicates position of the magnified panels (bottom). The solid white box indicates the position of the magnified panels of epidermis and generated epithelium (right). White scale bars, 3 mm; black scale bar, 100 μm. Similar findings were confirmed in three animals. **b**, Appearance of the ulcer before (top left) and after (bottom left) biopsy of generated epithelium 28 days after administration of DGTM-AAV in *Pdgfra*^{creER};*LSL*^{tdTomato} mice and immunohistochemical analysis (top middle) of biopsied generated epithelium. Intensity of the tdTomato signal in the biopsied sample is higher than in the chamber (see **a**) because fixation without the chamber allowed the sample to contract. Yellow arrows indicate the position of the biopsy. The white dotted outline indicates the position of the magnified panels (lower middle). The white solid outline indicates the position of the magnified panels of generated epithelium (far right). White scale bars, 3 mm; black scale bar, 100 μm. Similar findings were confirmed in five biopsies. **c**, Experimental design of flow cytometric analysis of generated epithelium. **d**, Representative appearance of an ulcer 18 days after the administration of DGTM-AAV in *Pdgfra*^{creER};*LSL*^{tdTomato} mice (left) and primary cultured cells on feeder (middle) and no feeder condition (right). Yellow arrows indicate generated epithelium subjected to flow cytometric analysis. White scale bar, 3 mm; yellow scale bars, 200 μm. Similar findings were confirmed in three animals. **e**, Flow cytometric analysis of three primary cultured cells from the surface of an ulcer 18 days after administration of DGTM-AAV in *Pdgfra*^{creER};*LSL*^{tdTomato} mice (day 18 iSEPs). Primary keratinocytes from the back skin of *Krt14*^{cre};*LSL*^{tdTomato} mice (top) and wild-type mice (bottom) were prepared for positive and negative controls, respectively. For gating strategy example, see Supplementary Fig. 2. **f**, Schematic of stepwise operations performed with representative images. A wound was created and treated with DGTM-AAVs. After the generation of epithelial tissue, the initial skin chamber was removed and replaced with a larger one (twice the area of the small chamber). After the large chamber was removed, early contraction of the generated epithelium was prevented by a rubber ring. White scale bars, 5 mm. **g**, Chronological changes in surface area of generated epithelia in the large chamber. Coloured lines represent animals treated with DGTM-AAV ($n = 10$). The black line represents animals without AAV ($n = 5$). **h**, Representative images showing the gross appearance of ulcers in large chambers at different time points. White scale bars, 5 mm. Similar findings were confirmed in ten animals for DGTM-AAV^{+} group and five animals for the no-virus group. **i**, Appearance (top left), stereoscopic analysis (bottom left), H&E staining (top right) and immunohistochemical analysis (bottom right) of skin and subcutaneous tissue including generated epithelium on day 93. Yellow arrows indicate the periphery of the generated epithelium. Dotted lines in the left panels indicate the approximate position of the histological sections shown on the right. Stereoscopic analysis revealed areas of generated tissues that are not clear by appearance. Dotted outlines indicate position of magnified panels. Scale bars, 5 mm. Similar findings were confirmed in all animals monitored over days 90–110 ($n = 10$). **j**, A representative H&E staining of liver tissue 3 months after the administration of DGTM-AAVs for induction of lineage conversion. Scale bar, 1.0 cm. Similar findings were confirmed in five animals. **k**, Complete blood count analysis after administration of AAV-DGTM ($n = 4$ for experimental and control groups). **l**, Blood chemistry analysis after administration of DGTM-AAVs ($n = 3$ for experimental and control groups). **k**, **l**, Data are mean ± s.d.; statistical differences are analysed with a two-sided Student's *t*-test. **m**, Schematic of the skin island in chamber assay performed with representative images (intact skin group). **n**, Schematic of wounding, epithelialization and the skin island in chamber assay performed with representative images (epithelialized skin group). **o**, Schematic of skin biopsy, primary culture, cell-sheet transplantation and the skin island in chamber assay performed with representative images (cell sheet group). **p**, Schematic of chamber attachment and DGTM-AAV administration, chamber detachment, and the skin island in chamber assay performed with representative images (generated epithelium group). **q**, Appearances of skin island in chamber assay performed for pure cell sheet transplanted areas. Skin, purely composed of a cell sheet disappeared with time within the ulcer. Similar findings were confirmed in two animals. With these findings, we allowed the transplants to contract for another 1–2 weeks before subjecting the cell sheets to the skin island assay in the cell sheet group. **r**, Chronological changes in surface area of epithelial tissues after skin island creation and chamber attachment. Mean of four groups (left top), mean (thick line) with respective animals (thin line) for intact skin (top middle, $n = 8$), epithelialized skin (top right, $n = 8$), cell sheet (bottom left, $n = 6$), and generated epithelium (bottom right, $n = 8$) groups. **s**, Appearances of skin island in animals with maximum and minimum epithelialized areas for each group on day 14 of the skin island in chamber assay. The abilities of generated epithelium to laterally expand within an ulcer were more variable than seen with controls. **m**–**q**, **s**, External diameter of chamber, 14.3 mm. **t**, H&E staining and immunohistochemical analyses (K10 and K13) of generated tissues by DGTM-AAV at different time points (day 18, $n = 18$; days 28–30, $n = 20$; days 90–11, $n = 10$). Samples collected from the same generated epithelial tissues by partial biopsy (days 28–30) and thorough histological investigation (days 90–110) are indicated with asterisks (bottom two rows of the middle and left columns). In these two animals, K13 expression was confirmed in a biopsied sample at day 28, but after 3 months, K13 expression had been extinguished. Results are summarized in Fig. 4b. **u**, H&E staining and immunohistochemical analysis of mouse fetal skin at different gestational ages. Mouse fetal skin was transiently positive for K13. Similar findings were confirmed in three embryos from two different mothers for each gestational age.


a
VEGF-AAV
DGTM-AAV
GFPNLS-AAV
PBS
× 5 animals each
Histological analysis
D0
D7
b
HE
CD31 DAPI
HE
CD31
c
20
15
10
5
0
Microvessel densities (%)
P=0.0031
P=0.0082
P=0.0105
VEGF-AAV
DGTM-AAV
GFPNLS-AAV
PBS
d
20
15
10
5
0
Microvessel densities (%)
VEGF-AAV
DGTM-AAV
GFPNLS-AAV
PBS
e
Chamber attachment
DGTM-AAVs
Assessment
D0
D7
D25
f
D0
D7
D25
g
5.0 × 10¹¹ GC
h
GFPNLS-AAV
±
Collagen
× 4 animals each
3 days
Stereoscopic observation
Deep fascia and above tissue inside chamber
Liver
DNA extraction
qPCR
i
Collagen (+)
Collagen (-)
Central portion of ulcer
Liver surface
j
10¹¹
10⁹
10⁷
10⁵
10³
AAV copies per tissue
P=0.70
Col(-)
Col(+)
10⁸
10⁶
10⁴
10²
1
10⁻¹
AAV copies per diploid (vg/dg)
P=0.002
Col(-)
Col(+)
k
10
1
10⁻¹
10⁻²
AAV copies per diploid (vg/dg)
P=0.0002
Col(-)
Col(+)
l
DGTM-AAVs with collagen gel
Protocol I
Protocol II
Protocol III
Protocol IV
D0
D4
D17
Rock inhibitor (10 mM) (30 ul/day)
FGF2 (50 ug/ml) (50 ul/day)
m
D0
D4
D5
D6
D7
D8
D9
D10
D11
D12
D13
D14
D15
D16
D17
Protocol I
Protocol II
Protocol III
Protocol IV

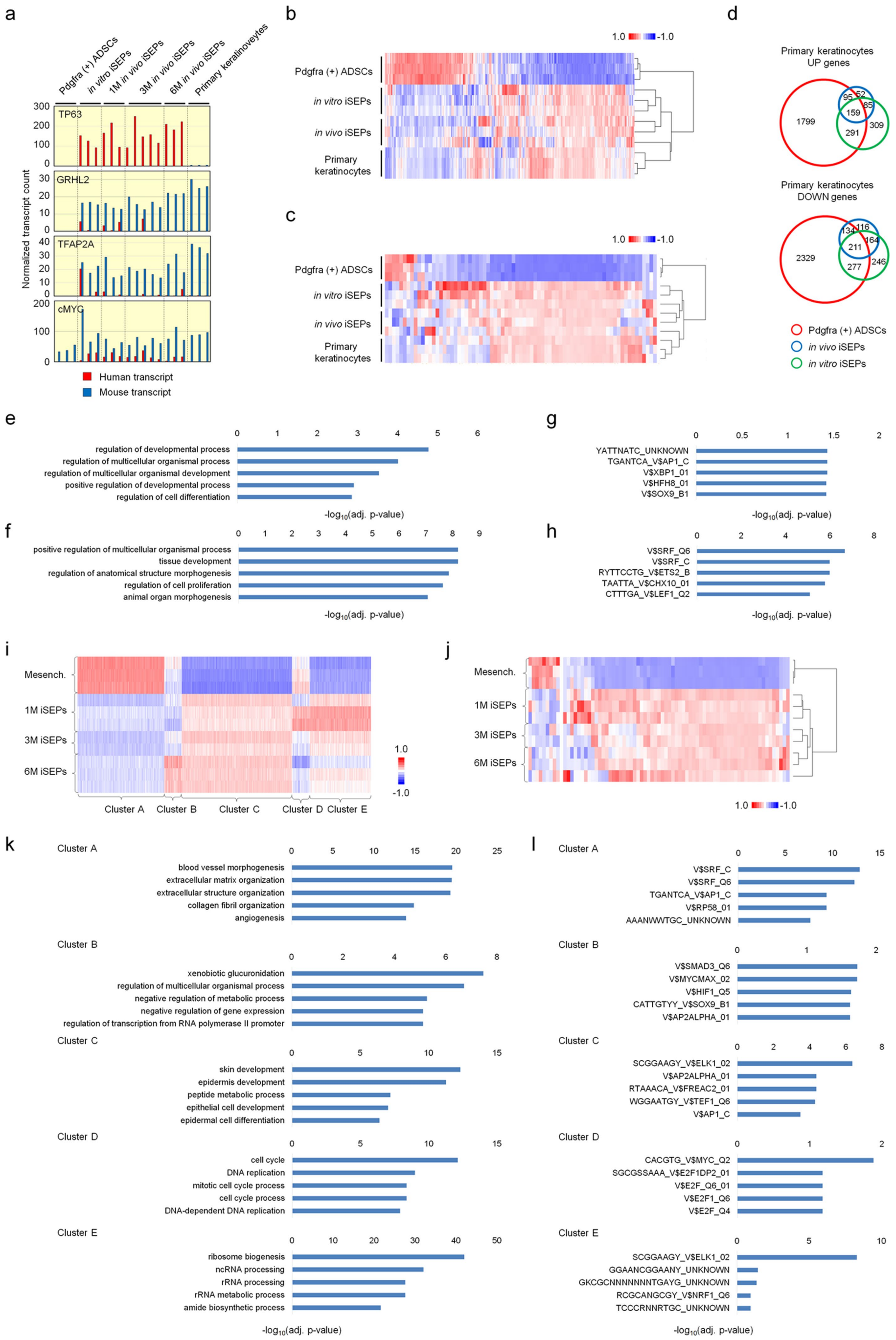
Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | In vivo reprogramming in a clinically relevant context. **a**, Schematic of investigation of influences of DGTM-AAVs administration on the vascularity of the wound bed. **b**, Appearance (top left) and immunohistochemical analysis of microvessel densities (right top and bottom) of an ulcer 7 days after the administration of DGTM-AAVs for CD31. Black and white boxes indicate the position of the magnified panels (lower). Yellow and red dotted lines indicate tissues above and below the deep fascia, respectively. Microvessel density was histologically evaluated for tissues above and below deep fascia. Scale bar, 1 mm. **c**, **d**, Analysis of microvessel densities for tissues above (**c**) and below (**d**) deep fascia in five animals for each group. Between six and eight sections were analysed for each animal. Mean microvessel densities are presented. Overlaid dot plots indicate the distribution of the data. Differences between groups were analysed with one-way ANOVA with a Tukey's multiple comparison test. DGTM-AAVs showed no influences on vascularity. **e**, Schematic of the experimental design for the investigation of the efficiency of in vivo reprogramming during the generation of epithelial tissues in old wounds (7 days after the creation of the ulcer). **f**, Representative appearances of an ulcer at each time point (on day of ulcer creation (day 0), upon administration of DGTM-AAVs (day 7) and 18 days after administration of DGTM-AAVs (day 25)). Scale bar, 3 mm. **g**, Appearances of ulcers 18 days after administration of DGTM-AAVs in ten animals (including one shown in **f**) and histological images of small generated epithelial tissues that are unidentifiable by appearance. Yellow arrows indicate visible generated epithelial tissues. Red arrows indicate the position of generated tissues in magnified panels. Red scale bars, 200 μm. **h**, Schematic of the experimental design for the investigation of efficiency of GFPNLS-AAV administration with or without collagen gel on ulcer. **i**, Stereoscopic analysis of the centre of an ulcer and the surface of the liver. Different exposure times were used for imaging of ulcer and liver (fluorescence of the ulcer is far stronger than that of the liver surface). Similar findings were confirmed in four animals for each group. **j**, qPCR analysis of gross AAV genomic copies in ulcer tissues (left) and AAV genomic copies per mouse diploid genome (right) in animals 3 days after administration of GFPNLS-AAV with or without collagen gel (four animals for each group). The displayed values are the minimum (bottom range), mean (holizontal line), and maximum (top range). Overlaid dot plots indicate the distribution of the data. **k**, qPCR analysis of AAV genome copies per mouse diploid genome in liver in animals 3 days after administration of GFPNLS-AAV with or without collagen gel (four animals for each group). Three different lobes of liver tissues were analysed for each animal. The displayed values are the minimum (bottom range), mean (holizontal line), and maximum (top range) for each animal. Overlaid dot plots indicate the distribution of the data. **j**, **k**, Differences between groups were analysed with two-sided Student's *t*-test. **l**, Schematic of the experimental design for epithelialization of ulcer in large chamber. **m**, Representative appearances of an ulcer treated by DGTM-AAVs and collagen gel administration (protocol I, $n = 2$, similar findings in both), an ulcer treated by DGTM-AAVs and collagen gel administration with application of Rock inhibitor (protocol II, $n = 2$, similar findings in both), an ulcer treated by DGTM-AAVs and collagen gel administration with application of FGF2 (protocol III, $n = 3$, similar findings in all), and an ulcer treated by DGTM-AAVs and collagen gel administration with application of Rock inhibitor and FGF2 (protocol IV, $n = 3$, similar findings in all). Yellow arrows indicate the initial emergence of visually identifiable epithelial tissues. External diameter of chamber, 14.3 mm.


a
NPKs
PKs
iSEPs
Single cell clone selection (7 Days/ 96 well)
Passaging
Expansion (7 Days/ 12 well)
Rhodanile blue Staining
b
PKs colonies (PK1)
iSEPs colonies (3M iSEPs 1)
c
PKs colonies (PK1)
iSEPs colonies (3M iSEPs 1)
d
%Area of colonies
100
50
0
NPK1 NPK2 NPK3 PK1 PK2 PK3 PK4 PK5 3M iS1 3M iS2 3M iS3 3M iS4 3M iS5 6M iS1 6M iS2 6M iS3 6M iS4 6M iS5
e
Hela
Primary keratinocytes
in vivo iSEPs
DMEM
DMEM KFM CKFM
DMEM KFM CKFM
f
Number of colonies (>30um)/ well
1,000
0
Hela DMEM KFM CKFM
PK1 PK2 PK3 PK4 PK5 3M iSEPs 1 3M iSEPs 2 3M iSEPs 3 3M iSEPs 4 3M iSEPs 5 6M iSEPs 1 6M iSEPs 2 6M iSEPs 3 6M iSEPs 4 6M iSEPs 5
g
(C57BL/6-TgN(act-EGFP)) mES
(stained with human MAB) Hela
(Krt14-Cre; LSL-tdTomato) PK
(Pdgfra-CreER;LSL-tdTomato) 3M iSEPs1 3M iSEPs2 3M iSEPs3
Phase
GFP
hMito DAPI
tdTomato
h
5.0 × 10⁶ cells + Matrigel (100ul) (200ul total)
mES
Hela
PK
3 iSEPs
Injection
×3 for each cells
Collection
D0
D28
i
Volume of nodule
300
200
100
0
650* 3468** 864
mES Hela PK 3M iSEPs1 3M iSEPs2 3M iSEPs3
j
mES Hela PK 3M iSEPs1 3M iSEPs2 3M iSEPs3
k
HE
tdTomato DAPI
l
HE
tdTomato DAPI
m
HE
GFP DAPI
n
HE
hMit DAPI
o
MSCV Luciferase PGK-hygro^r retrovirus
3M iSEPs 1
3M iSEPs 2
3M iSEPs 3
Hygromycin selection
2 week
Injection
D0
IVIS Collection
D56
p
3M iSEPs1 3M iSEPs2 3M iSEPs3
3M iSEPs1 3M iSEPs2 3M iSEPs3
q
3M iSEPs1
3M iSEPs2
3M iSEPs3
B Ey H Li Sp Lu K Pa
Es+St
SI
Ce
M
Co
Sk+N

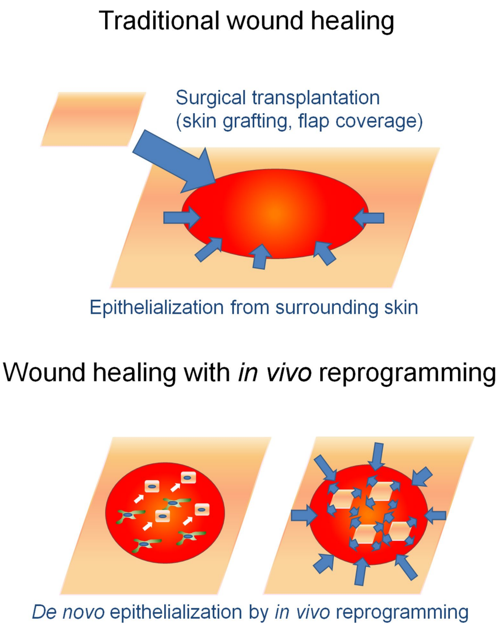
Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | in vivo iSEPs have high clonogenic ability and tumorigenic potential compatible with non-malignant epithelial cells. **a**, Schematic representation of clonogenicity assessment by expansion of single-cell-derived clones performed for neonatal primary keratinocytes (NPKs), primary keratinocytes from adult mice and in vivo iSEPs. **b**, Representative images of large (left), medium (middle), and small (right) single-cell-derived colonies of primary keratinocytes and iSEPs in 96-well plate. Similar findings were obtained from 122 (from three NPKs and five primary keratinocytes) and 169 (five 3M iSEPs and five 6M iSEPs) clones. **c**, Representative images of plates stained with Rhodanile blue staining. Similar differences between primary keratinocyte and 3M iSEPs were confirmed in five sets of primary keratinocytes and iSEPs isolated from the same animal. **d**, Occupied areas of single-cell-derived colonies in 12-well plates. iSEPs ($n = 10$) constantly show higher clonogenic ability than primary keratinocytes ($n = 8$). The displayed values are the minimum (bottom whisker), 25th percentile (bottom of box), median (line in box), 75th percentile (top of box), and maximum (top whisker). Overlaid dot plots indicate the distribution of the data ($n = 12$ for NPK1–3 and 6M iSEPs 1–5, $n = 15$ for PK1, $n = 21$ for PK2, $n = 18$ for PK3–4, $n = 14$ for PK5, $n = 18$ for 3M iSEPs 1, $n = 16$ for 3M iSEPs 2, $n = 29$ for 3M iSEPs 3, $n = 24$ for 3M iSEPs 4, and $n = 22$ for 3M iSEPs 5). **e**, Representative images of soft-agar assay. In vivo iSEPs showed no colony formation in DMEM (DMEM + 10% FBS), KFM and CKFM. Scale bar, 200 μm. Similar findings were confirmed in one HeLa, five primary keratinocytes, five 3M iSEPs, and five 6M iSEPs. **f**, Number of colonies (>30 μm) in one well (six-well plate) ($n = 3$, technical replicates). Averages of triplicates are shown. Overlaid dot plots indicate the distribution of the data. **g**, Bright field and fluorescence images showing the representative morphology of mES cells from C57BL/6-TfN(act-EGFP) mice, fixed HeLa cells, primary keratinocytes from *Krt14*^{cre};*LSL*^{tdTomato} mice, and in vivo iSEPs from *Pdgfra*^{creER};*LSL*^{tdTomato} mice after in vivo reprogramming with DGTM-AAV. hMito, human mitochondria. Scale bar, 200 μm. Similar findings were confirmed once for mES and HeLa, five times for primary keratinocytes, and ten times for in vivo iSEPs. **h**, Schematic representation of the tumorigenic/teratogenic assay by subcutaneous injection of the cells to immunodeficient mice. **i**, Volume of the nodules 28 days after injection. Top, volume of mES-cell-derived nodules. Asterisks indicate the day on which nodules were collected (*, day 24; **, day 16) owing to animal-welfare concerns about the size or properties of the nodule. **j**, Appearances of the largest nodules for HeLa, primary keratinocytes, and iSEPs. Similar findings were confirmed in three animals for each. Scale bars, 5 mm. **k**–**n**, H&E staining and fluorescence analysis of the largest nodule of 3M iSEPs 1 (**k**), primary keratinocytes (**l**), mES cells (**m**), and HeLa cells (**n**). For mES cells and HeLa cells teratogenicity and transfacial invasion were confirmed, respectively. Dotted outlines indicate the position of magnified panels (bottom). Similar findings were observed in other samples for each cells ($n = 9$ for iSEPs, $n = 3$ for primary keratinocytes, mES, and Hela cells). Red scale bar, 5 mm; blue scale bar, 2 mm; black scale bar, 500 μm; yellow scale bar, 200 μm. **o**, Schematic representation of biodistribution assessment of iSEPs. luciferase-expressing iSEPs were prepared using retroviral transduction. **p**, In vivo luminescent images 56 days after the subcutaneous injection of luciferase-expressing iSEPs. **q**, Luminescent images of dissected organs and tissues. Luciferase expression was confined to site of injection. M, muscle beneath the transplants; Sk + N, nodule and overlaid skin. **p**, **q**, Similar findings were confirmed in three animals for each iSEP group.



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Exogenous gene expression and transcriptional profiles of iSEPs. a, Integration analysis of *TP63*, *GRHL2*, *TFAP2A* and *MYC*. RNA-seq reads were mapped simultaneously to human and mouse transcriptomes and the sum of the normalized transcript counts for all variants of *TP63*, *GRHL2*, *TFAP2A*, and *MYC* is shown for Pdgfra$^+$ ADSCs, in vitro iSEPs, 1M in vivo iSEPs, 3M in vivo iSEPs, 6M in vivo iSEPs and primary keratinocytes. Human transcripts are derived from the AAV genome. Data from three Pdgfra$^+$ ADSCs, three in vitro iSEPs, three 1M in vivo iSEPs, five 3M in vivo iSEPs, six 6M in vivo iSEPs and three primary keratinocytes. **b**, Clustered heat map showing the normalized expression of the top expressed genes across all conditions. **c**, Clustered keratinocyte marker gene expression represented as a heat map. **d**, Differentially expressed genes were found between primary keratinocytes and Pdgfra$^+$ ADSCs (red), in vivo iSEPs (blue) and in vitro iSEPs (green). Overlap of genes that are significantly up- (top) or downregulated (bottom) in primary keratinocytes are shown. **e**, **f**, Top five Gene Ontology (GO) enrichment terms overrepresented for genes upregulated in primary keratinocytes (**e**) or genes upregulated in in vivo iSEPs (**f**). **g**, **h**, Top five enriched transcription factors known to bind to promoters of genes upregulated in primary keratinocytes (**g**) or genes upregulated in in vivo iSEPs (**h**). **e**–**h**, Bar plots show −log(adjusted *P* value) significance of overrepresentation. Enrichment testing was carried out using HOMER for Gene Ontology enrichment testing (hypergeometric test with Benjamini and Yekutieli general multiple testing correction), and WebGestalt for transcription-factor enrichment testing (hypergeometic test with Benjamini–Hochberg multiple testing correction) on genes upregulated in primary keratinocytes ($n = 399$), or upregulated in in vivo iSEPs ($n = 632$). **b**–**h**, Data for Pdgfra$^+$ mADSCs, in vitro iSEPs, 6M and 3M in vivo iSEPs were from three experiments. **i**, Normalized expression values were *k*-means clustered to reveal the top five patterns of gene expression. Cluster A, Mesench (mesenchymal); cluster B, late; cluster C, sustained; cluster D, early; cluster E, transient. **j**, Heat map showing relative normalized expression of keratinocyte marker genes. **k**, Analysis of the enrichment of the top Gene Ontology terms are shown for each cluster of genes. General terms with >100 genes were filtered out to remove overly general terms. **l**, Top five transcription factors enriched in promoters of clustered genes. **k**, **l**, Bar plots show –log(adjusted *P* value) significance of overrepresentation. Enrichment testing was carried out using HOMER for Gene Ontology enrichment testing (hypergeometric test with Benjamini and Yekutieli general multiple testing correction), and WebGestalt for transcription factor enrichment testing (hypergeometic test with Benjamini–Hochberg multiple testing correction) on genes in clusters A ($n = 1,966$), B ($n = 396$), C ($n = 2,501$), D ($n = 400$), and E ($n = 1,386$). **i**–**l**, Data are from three Pdgfra$^+$ mADSCs, three 1M in vivo iSEPs, two 3M in vivo iSEPs, three 6M in vivo iSEPs.



Traditional wound healing

Surgical transplantation
(skin grafting, flap coverage)
Epithelialization from surrounding skin

Wound healing with *in vivo* reprogramming

De novo epithelialization by *in vivo* reprogramming

Extended Data Fig. 10 | Description of wound healing with in vivo reprogramming. During physiological wound healing, epidermal defects are repaired from the other epidermis. On the other hand, in vivo reprogramming allows de novo epithelialization and greatly enhances the capacity for the regeneration of cutaneous defects.

LETTER

https://doi.org/10.1038/s41586-018-0471-x

Phosphocode-dependent functional dichotomy of a common co-receptor in plant signalling

Artemis Perraki[1,5], Thomas A. DeFalco[1,2], Paul Derbyshire[1], Julian Avila[3,4,6], David Séré[1,7], Jan Sklenar[1], Xingyun Qi[3,4], Lena Stransfeld[1,2], Benjamin Schwessinger[1,8], Yasuhiro Kadota[1,9], Alberto P. Macho[1,10], Shushu Jiang[1,11], Daniel Couto[1,12], Keiko U. Torii[3,4], Frank L. H. Menke[1] & Cyril Zipfel[1,2]*

Multicellular organisms use cell-surface receptor kinases to sense and process extracellular signals. Many plant receptor kinases are activated by the formation of ligand-induced complexes with shape-complementary co-receptors[1]. The best-characterized co-receptor is BRASSINOSTEROID INSENSITIVE 1-ASSOCIATED KINASE 1 (BAK1), which associates with numerous leucine-rich repeat receptor kinases (LRR-RKs) to control immunity, growth and development[2]. Here we report key regulatory events that control the function of BAK1 and, more generally, LRR-RKs. Through a combination of phosphoproteomics and targeted mutagenesis, we identified conserved phosphosites that are required for the immune function of BAK1 in *Arabidopsis thaliana*. Notably, these phosphosites are not required for BAK1-dependent brassinosteroid-regulated growth. In addition to revealing a critical role for the phosphorylation of the BAK1 C-terminal tail, we identified a conserved tyrosine phosphosite that may be required for the function of the majority of *Arabidopsis* LRR-RKs, and which separates them into two distinct functional classes based on the presence or absence of this tyrosine. Our results suggest a phosphocode-based dichotomy of BAK1 function in plant signalling, and provide insights into receptor kinase activation that have broad implications for our understanding of how plants respond to their changing environment.

Receptor kinases control all aspects of plant life, ranging from development to stress response, and depend on ligand-induced interaction with co-receptors for their activation[1]. In particular, plant LRR-RKs often form complexes with short, shape-complementary co-receptors of the SOMATIC-EMBRYOGENESIS RECEPTOR KINASE (SERK) family[2]. *Arabidopsis* SERK3 (also known as BAK1) is the best-characterized member of this family. It forms ligand-induced complexes with FLAGELLIN SENSING 2 (FLS2) and EF-TU RECEPTOR (EFR)—the pattern-recognition receptors for the bacterial-pathogen-associated molecular patterns flagellin (or the derived epitope flg22) and EF-Tu (or the derived epitope elf18), respectively—and is thus a regulator of antibacterial immunity[3–5]. BAK1 and SERK1 also act as co-receptors for BRASSINOSTEROID INSENSITIVE 1 (BRI1), which perceives brassinosteroids to regulate plant growth and development[6–8]. However, the molecular mechanisms underlying BAK1 activation, or its functional specificity in diverse signalling pathways, are largely unknown[1].

Previous knowledge of BAK1 phosphorylation has been gained largely from in vitro studies that identified phosphosites affecting overall BAK1 kinase activity[9–13], thus providing limited information about dynamic regulatory events in vivo. The specific identification of in vivo BAK1 phosphosites during immune signalling is impeded by the impaired function in immunity of epitope-tagged BAK1 variants[14], which could otherwise facilitate BAK1 enrichment before analysis by mass-spectrometry[9]. Furthermore, the involvement of BAK1 in several signalling pathways[2] makes it difficult to identify pathway-specific phosphosites from a total BAK1 cellular pool.

To gain insights into the mechanisms of BAK1 activation during immune signalling, we enriched endogenous BAK1 in complex with GFP-tagged FLS2 or EFR after ligand treatment (Extended Data Fig. 1a, b), and then used tandem mass spectrometry to identify phosphosites of native, immune-active BAK1. To reduce residual levels of brassinosteroid-activated BAK1, we pre-treated tissues with brassinazole (BRZ), an inhibitor of brassinosteroid biosynthesis. The analysis identified four BAK1 phosphosites (S602, T603, S604 and S612; see Extended Data Fig. 1c and Supplementary Fig. 2 for spectra) that were previously uncharacterized in vivo, in addition to T446, of which phosphorylation was found to be dispensible for both brassinosteroid and flg22 signalling[9]. Notably, S604 and S612 had been previously identified as in vitro phosphosites[10].

To test the function of these in vivo phosphosites, we generated stable transgenic *Arabidopsis* lines expressing non-phosphorylatable variants for these residues (substituting S or T with A) in the null *bak1-4* mutant[3] background, and tested for functional complementation by measuring flg22-induced production of reactive oxygen species (ROS) as an early immune output[15]. In the case of the S602, T603 and S604 sites we used triple A mutants (hereafter AAA), because single or double mutants did not impair *bak1-4* complementation in mesophyll protoplasts (Extended Data Fig. 2a, b) or stable lines (Extended Data Fig. 2c, d). Both the AAA and S612A mutants failed to complement flg22-induced ROS production in *bak1-4* (Fig. 1a, Extended Data Fig. 3a, b). By contrast, corresponding phosphomimetic (S602D/T603D/S604D or S612D) mutants partially complemented the impaired ROS phenotype of *bak1-4* protoplasts, in keeping with the importance of these residues as phosphosites, although these mutations were not gain-of-function (Extended Data Fig. 2e, f). Loss of S602/T603/S604 phosphorylation and loss of S612 phosphorylation also impaired both flg22-induced MAP kinase (particularly MPK4 and MPK11) activation (Fig. 1b) and immunity to the bacterial pathogen *Pseudomonas syringae* pv. *tomato* (*Pto*) DC3000, resembling results seen with the semi-dominant mutant *bak1-5* (ref. [16]) (Fig. 1c). This impaired immune function was not caused by impaired BAK1 accumulation or FLS2–BAK1 association (Fig. 1b, Extended Data Fig. 3c, d).

Importantly, the complementation mutant lines *bak1-4*/*BAK1*^{AAA} and *bak1-4*/*BAK1*^{S612A} did not show impaired response to endogenous brassinosteroid (Extended Data Fig. 4) or exogenously applied

---

[1]The Sainsbury Laboratory, Norwich Research Park, Norwich, UK. [2]Institute of Plant and Microbial Biology and Zürich-Basel Plant Science Center, University of Zürich, Zürich, Switzerland. [3]Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA. [4]Department of Biology, University of Washington, Seattle, WA, USA. [5]Present address: Department of Plant Sciences, University of Cambridge, Cambridge, UK. [6]Present address: Metabolomics Platform, The Broad Institute, Cambridge, MA, USA. [7]Present address: Laboratoire de Biochimie et Physiologie Moléculaire des Plantes, Institut de Biologie Intégrative des Plantes 'Claude Grignon', UMR CNRS/INRA/SupAgro/UM2, Montpellier, France. [8]Present address: The Australian National University, Research School of Biology, Acton, Australian Capital Territory, Australia. [9]Present address: RIKEN Center for Sustainable Resource Science, Yokohama, Japan. [10]Present address: Shanghai Center for Plant Stress Biology, CAS Center for Excellence in Molecular Plant Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China. [11]Present address: Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China. [12]Present address: Department of Botany and Plant Biology, University of Geneva, Geneva, Switzerland. *e-mail: cyril.zipfel@botinst.uzh.ch

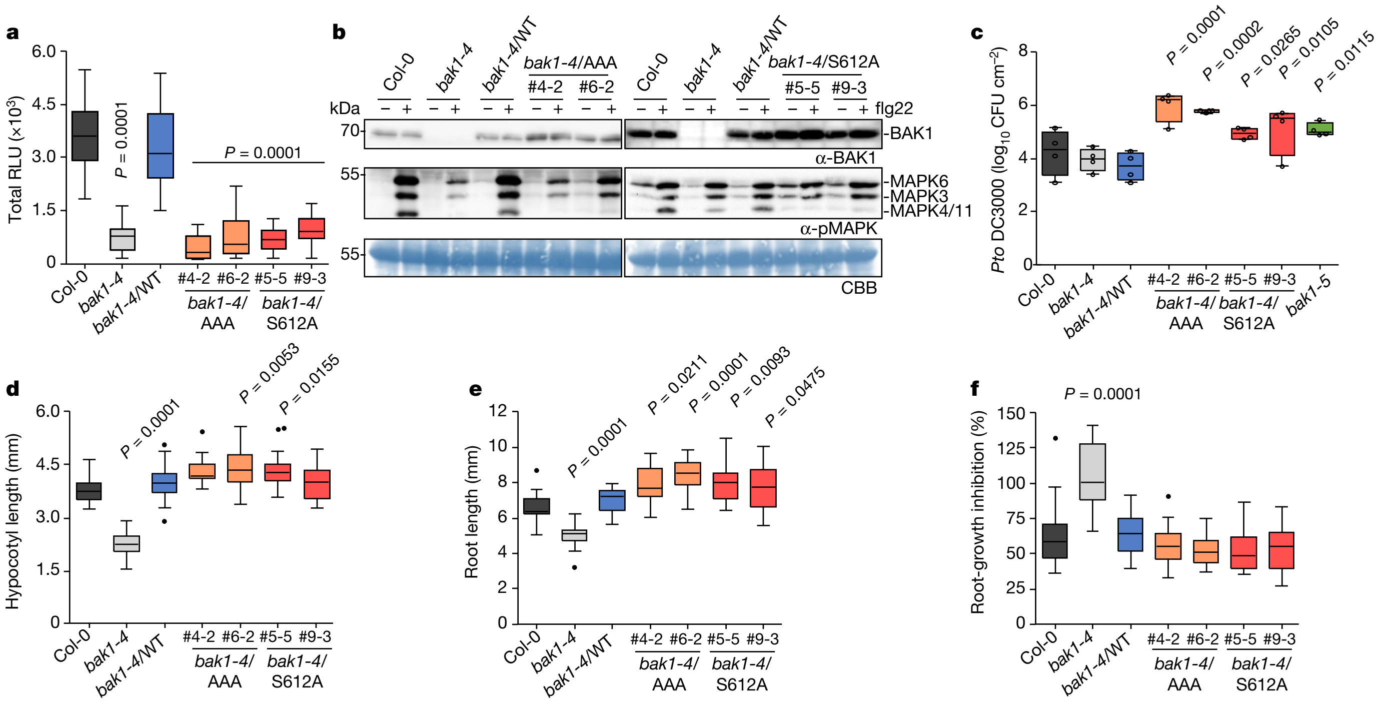


Fig. 1 | Phosphorylation of the BAK1 C-terminal tail is essential for immune signalling. a, Total ROS production following treatment of leaf discs with 100 nM flg22 over 60 min. $n = 12$ biologically independent discs. RLU, relative luminescence unit. **b**, MAP kinase activation 5 min after treatment of ten-day-old seedlings with 100 nM flg22. Coomassie Brilliant Blue (CBB) was used as loading control. For blot source data, see Supplementary Fig. 1. **c**, Growth of *Pto* DC3000 at 3 days post-inoculation. Plants were sprayed with a bacterial suspension containing 10^{8} colony-forming units (CFU) per ml. $n = 4$ biologically independent plants. Individual data points are shown as circles. **d**, Hypocotyl length of dark-grown seven-day-old seedlings grown on medium supplemented with 2 μM BRZ. Number of biologically independent hypocotyls: Col-0 ($n = 19$), *bak1-4* ($n = 27$), *bak1-4/BAK1*^{WT} (*bak1-4*/WT; $n = 28$), *bak1-4/BAK1*^{AAA}#4-2 (*bak1-4*/AAA; $n = 18$), *bak1-4/BAK1*^{AAA}#6-2 ($n = 22$), *bak1-4/BAK1*^{S612A}#5-5 (*bak1-4*/S612A; $n = 33$), *bak1-4/BAK1*^{S612A}#9-3 ($n = 29$) (numbers following hash symbols denote the independent lines used). **e**, Root length of light-grown seven-day-old seedlings grown on medium supplemented with 2 μM BRZ. Number of biologically independent roots: Col-0 ($n = 21$), *bak1-4* ($n = 21$), *bak1-4/BAK1*^{WT} ($n = 21$), *bak1-4/BAK1*^{AAA}#4-2 ($n = 25$), *bak1-4/BAK1*^{AAA}#6-2 ($n = 27$), *bak1-4/BAK1*^{S612A}#5-5 ($n = 24$), *bak1-4/BAK1*^{S612A}#9-3 ($n = 25$). **f**, Relative root-growth inhibition of light-grown seven-day-old seedlings treated with 5 nM brassinolide. Control and treated media were supplemented with 0.5 μM BRZ. Root-growth inhibition by brassinolide was calculated relative to the untreated control for each genotype for the following numbers of biologically independent roots: Col-0 ($n = 17$), *bak1-4* ($n = 23$), *bak1-4/BAK1*^{WT} ($n = 24$), *bak1-4/BAK1*^{AAA}#4-2 ($n = 25$), *bak1-4/BAK1*^{AAA}#6-2 ($n = 27$), *bak1-4/BAK1*^{S612A}#5-5 ($n = 22$), *bak1-4/BAK1*^{S612A}#9-3 ($n = 23$). For **a–f**, experiments were repeated independently three times. Measurements are plotted as box plots displaying the first and third quartiles, split by the median; whiskers extend to a maximum of 1.5× interquartile range beyond the box. Outliers are indicated as black dots. Statistical analysis was performed using a one-way ANOVA with Dunnett's post hoc test compared to *bak1-4/BAK1*^{WT}.

brassinolide (the most biologically active brassinosteroid), or BRZ (Fig. 1d–f, Extended Data Fig. 4e). This demonstrates that the conserved S602/T603/S604 and S612 phosphosites[17] (Extended Data Fig. 5a, b) are required for flg22-induced immune responses, but are dispensable for brassinosteroid signalling, and provides a mechanistic explanation for the phenotype caused by the deletion of the C-terminal tail of BAK1[18]. Using a phospho-specific antibody for S612 (Extended Data Fig. 5c, d), we observed that S612 phosphorylation is lost in C-terminally tagged BAK1 variants (Extended Data Fig. 5e), which may underlie their previously documented lack of functionality in immune signalling, although we cannot exclude the possibility of additional mechanism(s).

Plant receptor kinases are dual-specificity kinases, similar to animal receptor tyrosine kinases[19]; however, unbiased proteomics-based identification of phosphotyrosines in plant receptor kinases is not trivial[20–22]. We therefore also individually mutated tyrosine residues in the BAK1 cytoplasmic domain to non-phosphorylatable phenylalanine, and tested for complementation of elf18-induced ROS in *bak1-4* mesophyll protoplasts. The Y403F and Y463F mutants did not support ROS production, whereas it was partially impaired in the Y365F mutant (Extended Data Fig. 6a). The results of the Y463F mutant are consistent with previous findings that this residue is essential for the kinase activity of BAK1[13], therefore we did not pursue its characterization. However, considering the marked effect of the Y403F mutation on elf18-induced ROS production (Extended Data Fig. 6a), we generated *bak1-4/BAK1*^{Y403F} complementation mutant lines (Extended Data Fig. 7a, c). BAK1(Y403F) accumulated to wild-type levels and still formed a flg22-induced complex with FLS2 (Extended Data Fig. 7a, b), but its function in immunity was abrogated (Fig. 2a–c). As with the AAA and S612A mutants (Fig. 1d–f, Extended Data Fig. 4), the Y403F mutation did not impair brassinosteroid signalling (Fig. 2d–f, Extended Data Figs. 4e, 7c–e). Unlike phosphoserine and phosphothreonine, acidic residues are generally unable to mimic the charge and size of phosphotyrosine, which is consistent with the inability of the Y403D mutant to complement the elf18-induced ROS burst in *bak1-4* protoplasts (Extended Data Fig. 6b, c).

In a similar manner to the Y403F mutation, the semi-dominant *bak1-5* mutation affects immune but not brassinosteroid signalling, with the causative mutation C408Y being close to Y403 in the BAK1 catalytic loop[20] (Extended Data Fig. 8a). In silico substitution suggested that the mutation C408Y leads to steric hindrance with Y403 (Fig. 3a), which could affect the phosphorylation of Y403. To test this hypothesis, we generated C408Y, C408A, C408S or C408F BAK1 mutants, and assayed them for *bak1-4* complementation (Fig. 3a, b). Notably, only substitution with aromatic residues (C408Y and C408F) abolished BAK1 functionality (Fig. 3b), which suggests that the *bak1-5* phenotype is not caused by the generation of a new potential phosphosite (that is, C408Y) or by the loss of potential redox regulation on C408 (for example, C408A, C408S)[23]. Using a phospho-specific pY403 antibody (Extended Data Fig. 9a, b), we determined that the mutations C408Y

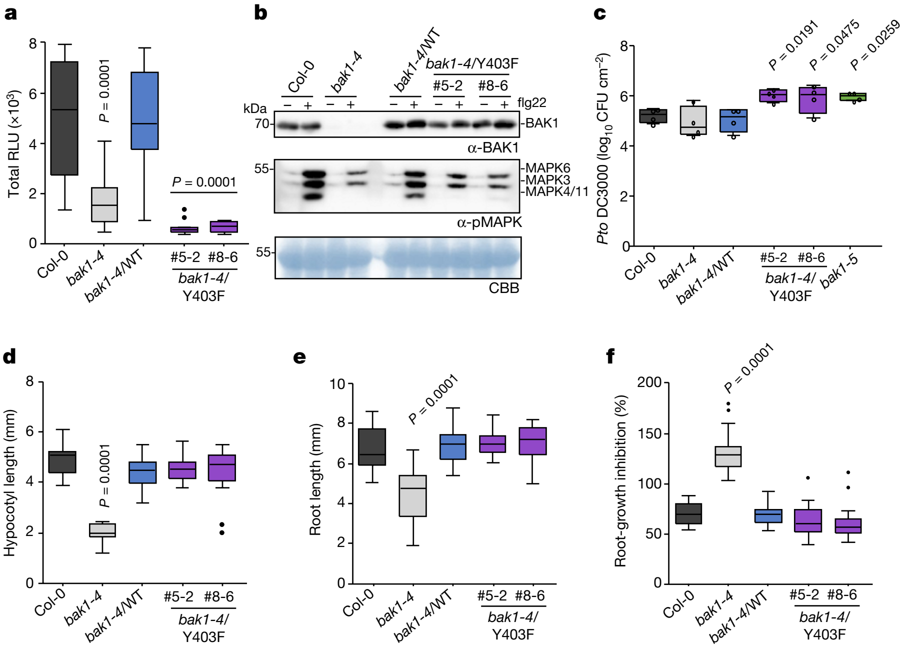


Fig. 2 | Phosphorylation of Y403 is required for BAK1 function in immune signalling. a, Total ROS production following treatment of leaf discs with 100 nM flg22 over 60 min. $n = 12$ biologically independent leaf discs. **b**, MAP kinase activation 5 min after treatment of ten-day-old seedlings with 100 nM flg22. CBB was used as loading control. For blot source data, see Supplementary Fig. 1. **c**, Growth of *Pto* DC3000 3 days post-inoculation. Plants were sprayed with a bacterial suspension containing 10^8 colony-forming units per ml. $n = 4$ biologically independent plants. Individual data points are shown as circles. **d**, Hypocotyl length of dark-grown seven-day-old seedlings grown on medium supplemented with 2 μM BRZ. Number of biologically independent hypocotyls: Col-0 ($n = 15$), *bak1-4* ($n = 13$), *bak1-4/BAK1*^{WT} ($n = 15$), *bak1-4/BAK1*^{Y403F}#5-2 (*bak1-4*/Y403F; $n = 22$), *bak1-4/BAK1*^{Y403F}#8-6 ($n = 22$). **e**, Root length of light-grown seven-day-old seedlings grown on medium supplemented with 2 μM BRZ. Biologically independent roots: Col-0 ($n = 17$), *bak1-4* ($n = 18$), *bak1-4/BAK1*^{WT} ($n = 17$), *bak1-4/BAK1*^{Y403F}#5-2 ($n = 20$), *bak1-4/BAK1*^{Y403F}#8-6 ($n = 23$). **f**, Relative root-growth inhibition of light-grown seven-day-old seedlings treated with 5 nM brassinolide. Control and treated media were supplemented with 0.5 μM BRZ. Root-growth inhibition by brassinolide was calculated relative to the untreated control for each genotype, for the following numbers of biologically independent roots: Col-0 ($n = 19$), *bak1-4* ($n = 23$), *bak1-4/BAK1*^{WT} ($n = 17$), *bak1-4/BAK1*^{Y403F}#5-2 ($n = 18$), *bak1-4/BAK1*^{Y403F}#8-6 ($n = 20$). For **a–f**, experiments were repeated independently three times. Measurements are plotted as box plots displaying the first and third quartiles, split by the median; whiskers extend to a maximum of 1.5× interquartile range beyond the box. Outliers are indicated as black dots. Statistical analysis was performed using a one-way ANOVA with Dunnett's post hoc test compared to *bak1-4/BAK1*^{WT}.

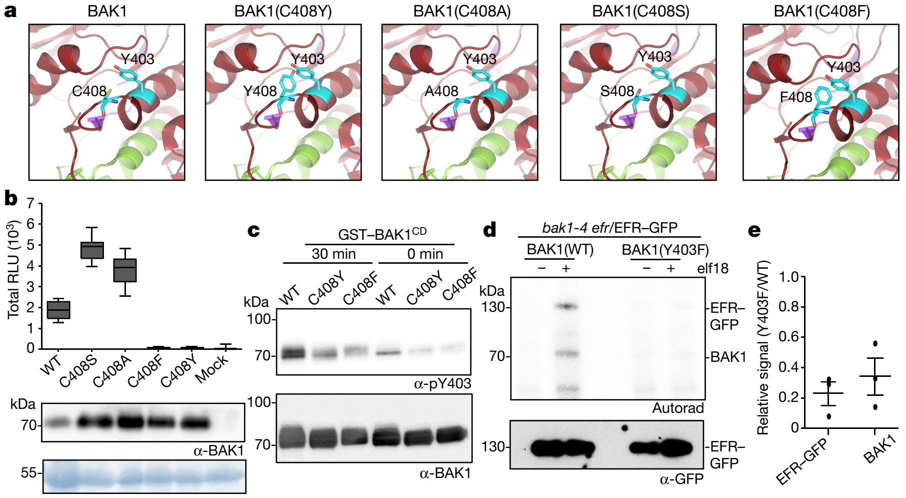


Fig. 3 | Phosphorylation of the conserved BAK1 residue Y403 is important for ligand-induced activation of the immune receptor complex. a, In silico substitutions of BAK1 C408 on a selected region of the BAK1 cytoplasmic domain structure (Protein Data Bank (PDB) ID: 3UIM). The activation segment (green) and catalytic loop (purple) are highlighted. **b**, Total ROS production of *bak1-4* mesophyll protoplasts transiently expressing the indicated BAK1 mutants, following treatment with 100 nM flg22 for 60 min. $n = 12$ biological independent suspensions. Measurements are plotted as box plots displaying the first and third quartiles, split by the median; whiskers extend to a maximum of 1.5× interquartile range beyond the box. **c**, In vitro phosphorylation of Y403 in recombinant proteins of the BAK1 cytoplasmic domain after a 30-min kinase reaction as detected by anti-pY403. Blots were further probed with anti-BAK1 as loading control. In **b** and **c**, experiments were repeated independently three times. **d**, Phosphorylation of EFR–GFP and associated BAK1 after treatment with water (−) or 100 nM elf18 (+) for 10 min. EFR–GFP was immunoprecipitated using GFP–Trap beads and then assessed by an in vitro kinase assay with [^{32}P]γ-ATP. EFR–GFP levels were determined by western-blot analysis with anti-GFP antibodies. For gel source data for **b–d**, see Supplementary Fig. 1. **e**, Aligned dot blots show the relative reductions in BAK1 and EFR–GFP phosphorylation levels in the *bak1-4 efr/BAK1*^{Y403F} EFR–GFP sample, as measured in three independent in vitro kinase assays as in **d**. Data are mean ± s.e.m.

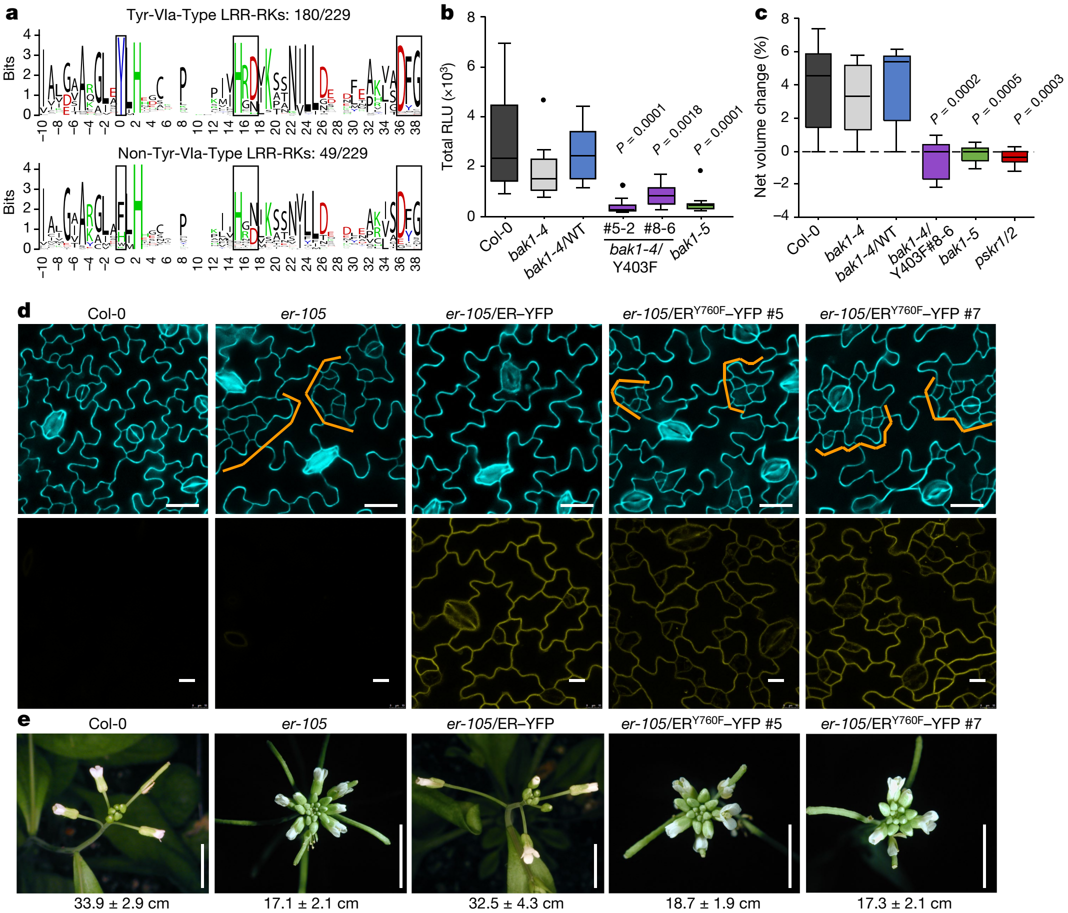


Fig. 4 | BAK1 Y403 and analogous tyrosine phosphorylation has a wide functional role in *Arabidopsis* LRR-RKs. a, WebLogo representation of the region analogous to BAK1 Y403 in *Arabidopsis* LRR-RKs. **b**, Total ROS production following treatment of leaf discs with 100 nM Pep1 for 60 min. $n = 12$ biologically independent discs. **c**, Cell expansion of mesophyll protoplasts (as measured by net volume change) following treatment with 1 nM PSK-α. $n = 10$ biologically independent protoplasts. In **b** and **c**, measurements are plotted as box plots displaying the first and third quartiles, split by the median; whiskers extend to a maximum of 1.5 × interquartile range beyond the box. Outliers are indicated as black dots. Statistical analysis was performed using a one-way ANOVA and Dunnett's post hoc test compared to *bak1-4/BAK1*^{WT}. **d**, Confocal microscopy images of abaxial cotyledon epidermis. Top, epidermal phenotypes from ten-day-old seedlings. Cell outlines are visualized with propidium iodide staining; bottom, ER–YFP expression from three-day-old seedlings. **e**, Representative images of inflorescence from four- to five-week-old plants. Scale bars, 5 mm. The main inflorescence lengths were measured, with mean ± s.e.m. indicated below the images. $n = 11$ (Col-0), $n = 18$ (*er-105*), $n = 8$ (*er-105/ER–YFP*), $n = 19$ (*er-105/ER*^{Y760F}–YFP #5) and $n = 19$ (*er-105/ER*^{Y760F}–YFP #7) biologically independent plants, at seven weeks old. In **b**–**e**, experiments were repeated at least twice with similar results.

and C408F reduce the phosphorylation of Y403 in vitro (Fig. 3c), suggesting that the *bak1-5* phenotype results from reduced Y403 phosphorylation, although we cannot fully exclude the possibility that additional mechanisms are also affected. Together with the finding that the Y403F mutation affects elf18-induced phosphorylation of the EFR complex[21] (Fig. 3d, e), these data suggest that Y403 phosphorylation is critical for the activation of immune signalling.

All BAK1 phosphosite mutants maintained both autophosphorylation (Extended Data Fig. 9c, d) and substrate *trans*-phosphorylation (Extended Data Fig. 9e) activities in vitro, although activity levels varied between mutants (Extended Data Fig. 9c–e). Whereas loss of BAK1 activity causes defects in brassinosteroid signalling[9], no quantitative defects were observed with these mutants (Figs. 1d–f, 2d–f). Phosphorylation at these sites is thus specifically associated with immune signalling and not solely the capacity of BAK1 as a kinase. Y403 and S612 are BAK1 autophosphorylation sites in vitro (Extended Data Fig. 9a, b) and, in the case of S612, autophosphorylation can occur *in trans* (Extended Data Fig. 9f).

Altogether, our data could suggest a phosphosite-dependent uncoupling of the functions of BAK1 in immunity and growth. However, upon close examination of *Arabidopsis* LRR-RK sequences, we noted that Y403 is strictly conserved in SERKs across plant species (Extended Data Fig. 8b), and that the analogous residue (which we denote Tyr-VIa, as it is present in kinase subdomain VIa) is conserved in around 80% of *Arabidopsis* LRR-RKs regardless of their function in growth or immunity (Fig. 4a, Extended Data Fig. 10a, Supplementary Table 1). Notably, phosphorylation of Tyr-VIa (Y836) is essential for EFR functionality, and the tyrosine phosphorylation of receptor kinases is actively inhibited by bacteria to cause disease[21].

We therefore postulated that the functionality of LRR-RKs—acting as the main ligand-binding receptors in diverse pathways—that contain this conserved Tyr-VIa residue would require phosphorylation of this residue as well as of the analogous residue in their BAK1 or SERK co-receptor(s). This hypothesis is so far supported by the fact that FLS2, EFR, the immune receptors PEP RECEPTOR 1 and PEP RECEPTOR 2 (PEPR1 and PEPR2), as well as ERECTA (ER) and the development-associated LRR-RKs HAESA/HAESA-LIKE 2 (HAE and HSL2) are *bak1-5*-sensitive[5,16,24,25]. By contrast, the Tyr-VIa position in BRI1 is occupied by a phenylalanine (Extended Data Fig. 10b), and brassinosteroid signalling is not affected by either *bak1-5* (ref. [16]) or by BAK1(Y403F) mutations (Fig. 2d–f, Extended Data Fig. 7c–e). We also show that the function of PEPR1 and PEPR2 requires BAK1 Y403 phosphorylation (Fig. 4b), and that phytosulfokine (PSK)-induced cell expansion mediated by PSK RECEPTOR 1 and PSK RECEPTOR

2 (PSKR1 and PSKR2)[26] is also *bak1-5*- and BAK1(Y403F)-sensitive (Fig. 4c). The functionality of ERECTA also requires the phosphorylation of its Tyr-VIa residue, as ER(Y760F) is unable to complement the increased stomatal-lineage divisions or reduced pedicel length phenotypes of the *er-105* mutant[27] (Fig. 4d, e).

Altogether, our study suggests a phosphocode-based regulation that is likely to be important for the function of the majority of *Arabidopsis* LRR-RKs. This work reveals a mechanism by which the common co-receptor BAK1 (and, by extension, related SERKs) differentially regulates at least two classes of ligand-binding LRR-RKs, named here as the Tyr-VIa-type and non-Tyr-VIa-type (Fig. 4a, Extended Data Fig. 10a, b, Supplementary Table 1). Notably, *Arabidopsis* CHITIN ELICITOR RECEPTOR KINASE 1 (CERK1) is a co-receptor for LysM-type receptor kinases and also relies on phosphorylation of a Tyr-VIa residue[28]. Using structural modelling and sequence comparisons, we found that an analogous Tyr-VIa residue is also conserved and phosphorylated in EPIDERMAL GROWTH FACTOR RECEPTOR (EGFR)[29], and is similarly conserved in several other animal tyrosine kinases (Extended Data Fig. 10c, d), which suggests that the tyrosine-based regulation we have uncovered is important for different classes of receptor kinases across kingdoms. Future structural work will be key to understanding the molecular basis of this regulation.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0471-x.

Received: 30 October 2017; Accepted: 17 July 2018;
Published online 3 September 2018.

1. Hohmann, U., Lau, K. & Hothorn, M. The structural basis of ligand perception and signal activation by receptor kinases. *Annu. Rev. Plant Biol.* **68**, 109–137 (2017).
2. Ma, X., Xu, G., He, P. & Shan, L. SERKing coreceptors for receptors. *Trends Plant Sci.* **21**, 1017–1033 (2016).
3. Chinchilla, D. et al. A flagellin-induced complex of the receptor FLS2 and BAK1 initiates plant defence. *Nature* **448**, 497–500 (2007).
4. Heese, A. et al. The receptor-like kinase SERK3/BAK1 is a central regulator of innate immunity in plants. *Proc. Natl Acad. Sci. USA* **104**, 12217–12222 (2007).
5. Roux, M. et al. The *Arabidopsis* leucine-rich repeat receptor-like kinases BAK1/SERK3 and BKK1/SERK4 are required for innate immunity to hemibiotrophic and biotrophic pathogens. *Plant Cell* **23**, 2440–2455 (2011).
6. Nam, K. H. & Li, J. BRI1/BAK1, a receptor kinase pair mediating brassinosteroid signaling. *Cell* **110**, 203–212 (2002).
7. Li, J. et al. BAK1, an *Arabidopsis* LRR receptor-like protein kinase, interacts with BRI1 and modulates brassinosteroid signaling. *Cell* **110**, 213–222 (2002).
8. Sun, Y. et al. Structure reveals that BAK1 as a co-receptor recognizes the BRI1-bound brassinolide. *Cell Res.* **23**, 1326–1329 (2013).
9. Wang, X. et al. Sequential transphosphorylation of the BRI1/BAK1 receptor kinase complex impacts early events in brassinosteroid signaling. *Dev. Cell* **15**, 220–235 (2008).
10. Karlova, R. et al. Identification of *in vitro* phosphorylation sites in the *Arabidopsis thaliana* somatic embryogenesis receptor-like kinases. *Proteomics* **9**, 368–379 (2009).
11. Yun, H. S. et al. Analysis of phosphorylation of the BRI1/BAK1 complex in *Arabidopsis* reveals amino acid residues critical for receptor formation and activation of BR signaling. *Mol. Cells* **27**, 183–190 (2009).
12. Wu, X. et al. Transphosphorylation of *E. coli* proteins during production of recombinant protein kinases provides a robust system to characterize kinase specificity. *Front. Plant Sci.* **3**, 262 (2012).
13. Yan, L. et al. Structural basis for the impact of phosphorylation on the activation of plant receptor-like kinase BAK1. *Cell Res.* **22**, 1304–1308 (2012).
14. Ntoukakis, V., Schwessinger, B., Segonzac, C. & Zipfel, C. Cautionary notes on the use of C-terminal BAK1 fusion proteins for functional studies. *Plant Cell* **23**, 3871–3878 (2011).
15. Yu, X., Feng, B., He, P. & Shan, L. From chaos to harmony: responses and signaling upon microbial pattern recognition. *Annu. Rev. Phytopathol.* **55**, 109–137 (2017).
16. Schwessinger, B. et al. Phosphorylation-dependent differential regulation of plant growth, cell death, and innate immunity by the regulatory receptor-like kinase BAK1. *PLoS Genet.* **7**, e1002046 (2011).
17. aan den Toorn, M., Albrecht, C. & de Vries, S. On the origin of SERKs: bioinformatics analysis of the somatic embryogenesis receptor kinases. *Mol. Plant* **8**, 762–782 (2015).
18. Wu, D., Liu, Y., Xu, F. & Zhang, Y. Differential requirement of BAK1 C-terminal tail in development and immunity. *J. Integr. Plant Biol.* **60**, 270–275 (2018).
19. Macho, A. P., Lozano-Durán, R. & Zipfel, C. Importance of tyrosine phosphorylation in receptor kinase complexes. *Trends Plant Sci.* **20**, 269–272 (2015).
20. Wang, Y. et al. Assessment of BAK1 activity in different plant receptor-like kinase complexes by quantitative profiling of phosphorylation patterns. *J. Proteomics* **108**, 484–493 (2014).
21. Macho, A. P. et al. A bacterial tyrosine phosphatase inhibits plant pattern recognition receptor activation. *Science* **343**, 1509–1512 (2014).
22. Mitra, S. K. et al. An autophosphorylation site database for leucine-rich repeat receptor-like kinases in *Arabidopsis thaliana*. *Plant J.* **82**, 1042–1060 (2015).
23. Bender, K. W. et al. Glutaredoxin AtGRXC2 catalyses inhibitory glutathionylation of *Arabidopsis* BRI1-associated receptor-like kinase 1 (BAK1) *in vitro*. *Biochem. J.* **467**, 399–413 (2015).
24. Meng, X. et al. Ligand-induced receptor-like kinase complex regulates floral organ abscission in *Arabidopsis*. *Cell Rep.* **14**, 1330–1338 (2016).
25. Meng, X. et al. Differential function of *Arabidopsis* SERK family receptor-like kinases in stomatal patterning. *Curr. Biol.* **25**, 2361–2372 (2015).
26. Stührwohldt, N., Dahlke, R. I., Steffens, B., Johnson, A. & Sauter, M. Phytosulfokine-α controls hypocotyl length and cell expansion in *Arabidopsis thaliana* through phytosulfokine receptor 1. *PLoS ONE* **6**, e21054 (2011).
27. Shpak, E. D., Berthiaume, C. T., Hill, E. J. & Torii, K. U. Synergistic interaction of three ERECTA-family receptor-like kinases controls *Arabidopsis* organ growth and flower development by promoting cell proliferation. *Development* **131**, 1491–1501 (2004).
28. Suzuki, M. et al. Autophosphorylation of specific threonine and tyrosine residues in *Arabidopsis* CERK1 is essential for the activation of chitin-induced immune signaling. *Plant Cell Physiol.* **57**, 2312–2322 (2016).
29. Zhang, G. et al. Mass spectrometry mapping of epidermal growth factor receptor phosphorylation related to oncogenic mutations and tyrosine kinase inhibitor sensitivity. *J. Proteome Res.* **10**, 305–319 (2011).

Acknowledgements We thank K. Morehouse, M. Smoker, J. Taylor and the John Innes Centre Horticultural Services for technical help; V. Ntoukakis, M. Stegmann and J. Dang for technical advice; S. Huber, K. Bender and R. Zielinski for providing materials; Y. Belkhadir and S. Huber for reading the manuscript; and all members of the Zipfel laboratory for discussion. This work was supported by the Gatsby Charitable Foundation and the European Research Council (grant 'PHOSPHinnATE') (C.Z.); the Gordon and Betty Moore Foundation (grant GBMF3035) and the Howard Hughes Medical Institute (K.U.T.); the European Molecular Biology Organization (EMBO-LTFs 100-2017 to T.A.D. and 225-2015 to S.J.); the RIKEN Special Postdoctoral Research Fellowship, the JSPS Excellent Young Researcher Overseas Visit Program, and the Uehara Memorial Foundation (Y.K.); and the JIC/TSL PhD Rotation Program (B.S.).

Reviewer information *Nature* thanks S. Clouse and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions A.P. designed and conceived experiments, collected most of the data and wrote the manuscript. T.A.D. developed and collected data in many of the biochemical assays and participated in the preparation of the final manuscript. P.D. and J.S. performed proteomics-based analysis, under the supervision of F.L.H.M. J.A. and X.Q. designed and performed ERECTA-related experiments under the supervision of K.U.T. D.S. performed the protoplast swelling experiments. B.S., Y.K. and A.P.M. performed preliminary work on BAK1 phosphorylation. L.S., S.J. and D.C. provided technical help with experiments. C.Z. designed and conceived experiments, supervised the study and wrote the manuscript. All authors commented and agreed on the manuscript before submission.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0471-x.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0471-x.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to C.Z.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Plant material and growth conditions. All plants used in this study were in the *Arabidopsis thaliana* accession Col-0 background. The following mutants have been previously reported: *bak1-4* (ref. [3]), *bak1-5*, *bri1-301 bak1-4* (ref. [16]), *bak1-4*/pBAK1:BAK1 (ref. [30]), *bak1-4 bkk1-1*/pBAK1:BAK1–Flag, *bak1-4*/pBAK1:BAK1–HA3, *bak1-4*/pBAK1:BAK1–eGFP (ref. [14]), Col-0/pFLS2:FLS2–GFP (ref. [31]), *efr-1*/pEFR:EFR–GFP (ref. [32]), *er-105* (ref. [27]) and *er-105*/pER:ER–YFP (ref. [33]). The ER(Y760F)–YFP construct was introduced to the *er-105* background via *Agrobacterium*-mediated transformation using the floral dip method[34]. BAK1 wild-type and BAK1 phosphomutant constructs were similarly inserted to the *bak1-4* and *bri1-301 bak1-4* backgrounds via *Agrobacterium*-mediated transformation. Plants were grown on soil as one plant per pot in controlled rooms maintained at 20–22 °C with a 10 h photoperiod or in vitro in Murashige and Skoog (MS) medium supplemented with vitamins and 1% sucrose (Duchefa) with a 16 h photoperiod (at 22 °C). Physiological assays were performed at 4–5 weeks on soil-grown plants or at 14 days post germination on plants grown in vitro, unless otherwise specified. The *bak1-4*/*efr-1*/pEFR:EFR–GFP and lines were generated by genetic crossing. The *bak1-4*/*efr-1*/pEFR:EFR–GFP line was further crossed to *bak1-4*/pBAK1:BAK1 or *bak1-4*/pBAK1:BAK1(Y403F) lines to generate the F$_1$ individuals used in immunoprecipitation-kinase assays.

Molecular cloning. The coding sequence of BAK1 was inserted via BsaI digestion and ligation to the Golden-Gate-compatible vector pICSL22001 (generated by The Sainsbury Laboratory Synthetic Biology group) for expression in protoplasts under the 35S promoter. The BAK1 single and multiple mutants were initially generated by BsaI digestion and ligation of respective PCR fragments to the pICSL22001 vector. The BAK1 phosphorylation mutants were amplified by PCR from equivalent pICSL22001 plasmids and inserted using In-Fusion cloning (Clontech Laboratories) between BsrgI and BamHI restriction sites to an epiGreenB2 vector containing the whole genomic region of BAK1, with a 1,445-bp promoter fragment[16,30]. The resulting constructs were transformed into *Agrobacterium tumefaciens* strain Agl1 containing the pSOUP helper plasmid. PCR-amplified fragments of the cytoplasmic domain of BAK1 (BAK1-CD, amino acids 255–615)[35] and phosphorylation mutants were introduced into pOPINM vector (N-terminal 6×His-MBP fusion) using In-Fusion enzyme (Clontech Laboratories) for recombinant protein expression. The GST-BAK1 and GST-BAK1-5 are in the pGEX-4T1 vector and have been previously described[16]. The kinase-dead (D416N) mutation in the pOPINM vector and C408F mutation in the pGEX-4T1 vector (N-terminal GST fusion) were generated using a DpnI-mediated site-directed-mutagenesis protocol. The D416N mutation was also introduced into a pET28a (+):BAK1-CD (amino acids 250–615) clone that was provided by K. Bender and R. Zielinski (University of Illinois) using DpnI-mediated mutagenesis. The GST-BIK1* (K105E) clone in pGEX-4T1 has been described previously[36]. To express ERECTA Y760F in plants, pJM284 (pER:ER–YFP)[33] was used as a template for site-directed mutagenesis and was further transformed into the *Agrobacterium tumefaciens* strain GV3101/pMP90. The sequences of all primers used in this study are listed in Supplementary Table 2.

Chemicals. Synthetic flg22, elf18 and AtPep1 peptides were purchased from EZBiolab. Peptide sequences for flg22, elf18 and AtPep1 have been previously described[36–38]. Synthetic PSK-α peptide[26] was purchased from PeptaNova GmbH. All peptides were dissolved in sterile water. EpiBL was purchased from Xiamen Topusing Chemical and prepared as 20 mM stock solution in ethanol. BRZ was purchased from Sigma and prepared as 10 mM stock solution in DMSO.

Immunity-related bioassays. ROS burst, MAPK activation and bacteria spray infection assays were performed as previously described[30,39].

Protoplast assays. The isolation and transfections of *Arabidopsis* mesophyll protoplasts were performed as previously described[40], and the protocol was further optimized to measure ROS burst. In brief, following the plasmid-transfection and washing step, the protoplasts were resuspended in a W5 buffer containing 154 mM NaCl, 125 mM CaCl$_2$, 5 mM KCl and 2 mM MES adjusted with 2 mM Tris to pH 8, and directly distributed to a 96-well plate (100 μl of protoplast suspension per well). The protoplasts were left to rest for 16 h before treatment. Protoplasts were gently treated with luminol L-O12 (Wako Chemicals) (4 μM final concentration), HRP (10 μg ml^{-1} final concentration) and the indicated elicitor peptide in the same W5 pH 8.0 solution (final volume 200 μl per well). In the case of flg22-induced ROS, protoplasts were co-transfected with a 35S:FLS2-Flag construct[41]. Luminescence was recorded over a 40–60-min period using a charge-coupled device camera (Photek Ltd).

For PSK-induced protoplast swelling assays, *Arabidopsis* mesophyll protoplasts were isolated by digestion of cell walls as described previously[40], resuspended in a W5 buffer containing 2 mM MES 5.7, 154 mM NaCl, 125 mM CaCl2 and 10 mM KCl[26,42] and kept in the dark before treatment. Protoplasts were then treated with 1 nM PSK-α solution and observed under 10× magnification on a Leica DMB5500B fluorescent microscope. Images were acquired every 5 min over a period of 30 min. ImageJ software was used to measure the circumference of the protoplasts from which the net volume change was calculated.

Brassinosteroid-sensitivity assay. Seeds were surface-sterilized and individually placed in line on square Petri dishes containing 1/2 MS 1% sucrose, 0.8% phytoagar and 2 μM BRZ. The plates were placed at 4 °C for 2 days and then placed vertically in a growth chamber for 6 days. Pictures of the plates were then taken to measure root lengths were measured using ImageJ software. For the root growth inhibition assay, 5 nM of brassinolide was added to a similar medium containing 0.5 μM BRZ. For hypocotyl growth in the dark assay, plates were moved from at 4 °C to light for 3 h to stimulate germination, before wrapping them in aluminium foil and then placing them vertically in the growth room for 6 days. For BES1 dephosphorylation assays, seedlings were germinated on MS-agar supplemented with 1% sucrose for 5 days before transplanting to 6-well plates (five seedlings per well). One day before assays, seedlings were treated with 2 μM BRZ (final concentration). Twelve-day-old seedlings were treated with brassinolide or ethanol (mock) at the concentrations indicated for 60 min.

Western blot. Western blot analysis was performed with antibodies diluted in blocking solution (5% non-fat milk in TBS with 0.1% (v/v) Tween) at the following dilutions: anti-GFP–HRP (B-2, Santa Cruz), 1:5,000; anti-MBP–HRP (NEB), 1:5,000; anti-pERK 44/42 (ERK1-Thr202/Tyr204, Cell Signaling Technology), 1:3,000; anti-FLS2, 1:1,000[4]; anti-BAK1, 1:5,000[5]. The polyclonal anti-pY403 and anti-pS612 antibodies were produced by Abmart and raised against ARGLA(pY)LHDHC and C-QIENEYP(pS)GPR epitope peptides respectively and were used in 1:1,000 dilution. The anti-BES1 antibodies were produced by GenScript against the (C)EDLELTLGNGKAHS epitope peptide and were provided as a gift from S. Huber (University of Illinois), and were used in 1:1,000 dilution. In the case of immunoblotting of in vitro samples, blots were blocked with TBS with 0.1% (v/v) Tween, 5% gelatin from cold water fish skin (Sigma), and probed with the following antibodies in the same blocking solution: anti-pTyr-HRP (PY99, Santa Cruz), 1:2,000; anti-pThr (P-Thr-Polyclonal, Cell Signaling), 1:1,000; anti-BAK1, 1:5,000; anti-pY403, 1:1,000; anti-pS612, 1:3,000. The following secondary antibodies were used: anti-rabbit IgG-HRP Trueblot (Rockland, 18-8816-31, dilution 1:10,000) for detection of BAK1-FLS2 co-immunoprecipitation or anti-rabbit IgG (whole molecule)–HRP (A0545, Sigma, dilution 1:10,000) for all other western blots.

Immunoprecipitation and protein purification. FLS2-BAK1 immunoprecipitation was performed as previously described[14]. EFR–GFP immunoprecipitations coupled with kinase assays were performed as previously described[21]. Large-scale EFR–GFP and FLS2–GFP immunoprecipitations were performed for in vivo discovery of BAK1 phosphorylation sites. *Arabidopsis* Col-0/pFLS2:FLS2-GFP and *efr-1*/pEFR:EFR-GFP seedlings were grown in MS medium, 1% sucrose (plus vitamins) (Duchefa) in 6-well plates (five seedlings per well) for two weeks. One day before treatment, 100 nM BRZ (final concentration) was added to the medium. About 30 g of tissue per condition was treated with 1 μM elf18, flg22 or MS alone as mock control for 10 min (including 4 min of vacuum infiltration) and frozen in liquid nitrogen. Tissue was ground with liquid nitrogen to a fine powder and resuspended in 2 ml g^{-1} powder of extraction buffer (150 mM Tris-HCl, pH 7.5; 150 mM NaCl; 5 mM EDTA; 1% Igepal; 1% (v/v) protease inhibitor mixture, phosphatase inhibitor 2 and 3 (Sigma), 1 mM Na$_3$VO$_4$). Immunoprecipitation was performed with GFP–Trap beads, as previously reported[21]. Immunoprecipitates were eluted in SDS loading buffer and proteins were further analysed by SDS–PAGE and western blot.

6×His-MBP- or GST-tagged proteins were purified by Ni^{2+} affinity chromatography using Ni Sepharose High Performance (GE Healthcare) or GST-Bind (Millipore) resin, respectively. Following elution, proteins were immediately concentrated and washed using Amicon Ultra-4 Centrifugal Filter Concentrators (MWCO 4,000, Millipore) and stored in buffer (25 mM Tris-Cl, 10% glycerol, 1 mM dithiothreitol, pH 7.5) until use in assays. For gel and blot source data, see Supplementary Fig. 1.

Kinase assays. Unlabelled in vitro kinase assays were performed as previously described[35], except that 10 μM unlabelled ATP was used and specific phosphorylation was detected with the phospho-specific antibodies as described in the figure captions. To detect autophosphorylation *in trans*, assays were performed in the same manner but with 100 μM unlabelled ATP for 60 min. In vitro kinase assays with labelled [^{32}P]γ-ATP and EFR–GFP immunoprecipitation-coupled kinase assays were performed as previously described[21]. Quantification of ^{32}P autoradiography bands was performed using the ImageJ software. For gel and blot source data, see Supplementary Fig. 1.

Mass spectrometry analysis. Proteins were separated by SDS–PAGE (NuPAGE, Invitrogen), and after staining with SimplyBlue stain (Invitrogen), bands corresponding to BAK1 were excised and digested by trypsin and AspN. Liquid chromatography coupled with tandem mass spectrometry (LC–MS/MS) analysis was performed with an Orbitrap Fusion Trihybrid mass spectrometer (Thermo

Scientific) and a nanoflow-HPLC system (Dionex Ultimate 3000, Thermo Scientific) as described previously[43]. The peptide identification was performed by searching the *Arabidopsis* database (TAIR10) using Mascot (v.2.4.1 Matrix Science) as described previously[43], with the modification of allowing trypsin and AspN peptide termini. Scaffold (v4; Proteome Software) was used to validate MS/MS-based peptide and protein identifications and annotate spectra. The position of the modified residue and the quality of spectra for individual phosphopeptides were manually inspected and validated.

Confocal microscopy. The epidermal phenotypes were observed using confocal microscopes, Zeiss LSM700 (for cell periphery stained with propidium iodide (Thermo Fisher Scientific)) and Leica SP5-WLL (for YFP fluorescence) as described previously[44].

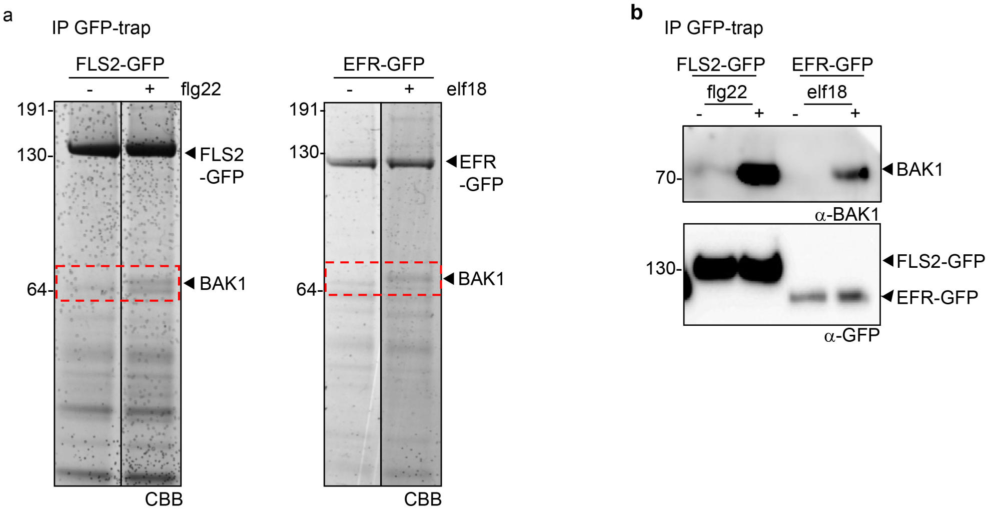
In silico analysis. Structural comparisons were performed using the DALI server[45]. The structure figures were prepared using PyMOL molecular viewer (http://www.pymol.org/). *Arabidopsis* LRR-RLKs were identified on the basis of previous reports[46,47], and protein sequences corresponding to the representative gene model of each were downloaded from TAIR10. Cytoplasmic receptor kinases (lacking a clear ectodomain) were removed, resulting in a final list of 229 LRR-RLKs, presented in Supplementary Table 1. These LRR-RKs were arranged by gene ID within previously assigned subgroups, and cytoplasmic domains were aligned by Clustal Omega to generate the WebLogos and alignments shown in Fig. 4a and Extended Data Fig. 9, respectively.

Statistical analysis. Statistical analysis was performed using one-way analysis of variance (ANOVA) and Dunnett's post hoc test, as implemented in GraphPad Prism 7.0. (GraphPad Software, http://www.graphpad.com).

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability. The data supporting the findings of this study are available within the paper and its Supplementary Information files. Source Data (gels and graphs) for Figs. 1–4 and Extended Data Figs. 1–7 and 9 are provided with the paper.

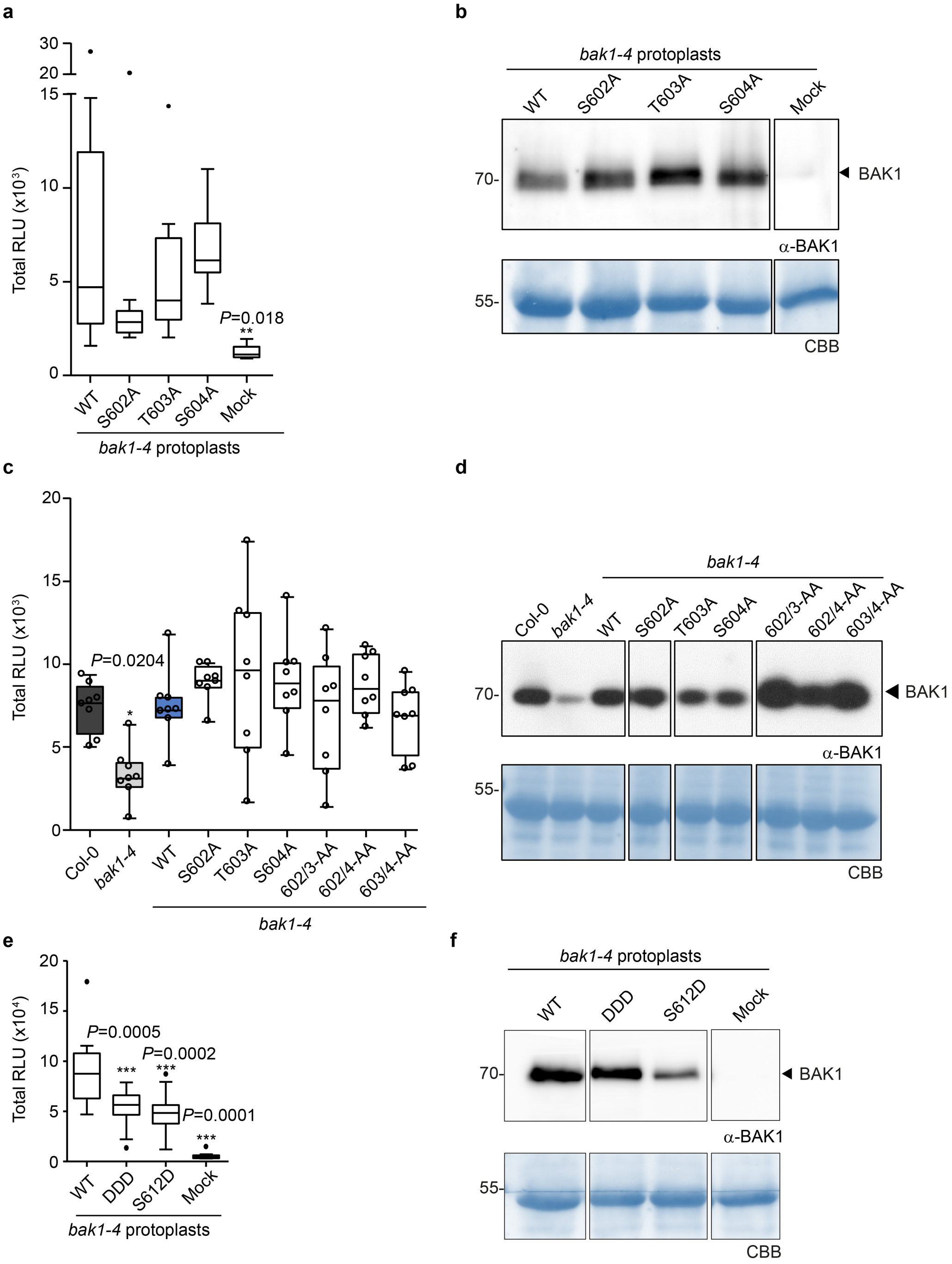
30. Singh, V. et al. Tyrosine-610 in the receptor kinase BAK1 does not play a major role in brassinosteroid signaling or innate immunity. *Front. Plant Sci*. **8**, 1273 (2017).
31. Robatzek, S., Chinchilla, D. & Boller, T. Ligand-induced endocytosis of the pattern recognition receptor FLS2 in *Arabidopsis*. *Genes Dev*. **20**, 537–542 (2006).
32. Nekrasov, V. et al. Control of the pattern-recognition receptor EFR by an ER protein complex in plant immunity. *EMBO J*. **28**, 3428–3438 (2009).
33. Lee, J. S. et al. Direct interaction of ligand–receptor pairs specifying stomatal patterning. Genes Dev. 26, 126–136 (2012).
34. Clough, S. J. & Bent, A. F. Floral dip: a simplified method for *Agrobacterium*-mediated transformation of *Arabidopsis thaliana*. *Plant J*. **16**, 735–743 (1998).
35. Kadota, Y. et al. Direct regulation of the NADPH oxidase RBOHD by the PRR-associated kinase BIK1 during plant immunity. *Mol. Cell* **54**, 43–55 (2014).
36. Felix, G., Duran, J. D., Volko, S. & Boller, T. Plants have a sensitive perception system for the most conserved domain of bacterial flagellin. *Plant J*. **18**, 265–276 (1999).
37. Huffaker, A., Pearce, G. & Ryan, C. A. An endogenous peptide signal in *Arabidopsis* activates components of the innate immune response. *Proc. Natl Acad. Sci. USA* **103**, 10098–10103 (2006).
38. Kunze, G. et al. The N terminus of bacterial elongation factor Tu elicits innate immunity in *Arabidopsis* plants. *Plant Cell* **16**, 3496–3507 (2004).
39. Monaghan, J. et al. The calcium-dependent protein kinase CPK28 buffers plant immunity and regulates BIK1 turnover. *Cell Host Microbe* **16**, 605–615 (2014).
40. Yoo, S.-D., Cho, Y.-H. & Sheen, J. *Arabidopsis* mesophyll protoplasts: a versatile cell system for transient gene expression analysis. *Nat. Protoc*. **2**, 1565–1572 (2007).
41. Sun, Y. et al. Structural basis for flg22-induced activation of the *Arabidopsis* FLS2-BAK1 immune complex. *Science* **342**, 624–628 (2013).
42. Ladwig, F. et al. Phytosulfokine regulates growth in *Arabidopsis* through a response module at the plasma membrane that includes CYCLIC NUCLEOTIDE-GATED CHANNEL17, H^+-ATPase, and BAK1. *Plant Cell* **27**, 1718–1729 (2015).
43. Bender, K. W. et al. Autophosphorylation-based calcium (Ca^{2+}) sensitivity priming and Ca^{2+}/calmodulin inhibition of *Arabidopsis thaliana* Ca^{2+}-dependent protein kinase 28 (CPK28). *J. Biol. Chem*. **292**, 3988–4002 (2017).
44. Horst, R. J. et al. Molecular framework of a regulatory circuit initiating two-dimensional spatial patterning of stomatal lineage. *PLoS Genet*. **11**, e1005374 (2015).
45. Holm, L. & Laakso, L. M. Dali server update. *Nucleic Acids Res*. **44**, W351–W355 (2016).
46. Lehti-Shiu, M. D. & Shiu, S.-H. Diversity, classification and function of the plant protein kinase superfamily. *Philos. Trans. R. Soc. Lond. B* **367**, 2619–2639 (2012).
47. Shiu, S. H. & Bleecker, A. B. Receptor-like kinases from *Arabidopsis* form a monophyletic gene family related to animal receptor kinases. *Proc. Natl Acad. Sci. USA* **98**, 10763–10768 (2001).


a
IP GFP-trap
FLS2-GFP
-
+
flg22
191-
130-
64-
◀FLS2
-GFP
◀BAK1
CBB
EFR-GFP
-
+
elf18
191-
130-
64-
◀EFR
-GFP
◀BAK1
CBB
b
IP GFP-trap
FLS2-GFP
flg22
EFR-GFP
elf18
-
+
-
+
70-
◀BAK1
α-BAK1
130-
◀FLS2-GFP
◀EFR-GFP
α-GFP

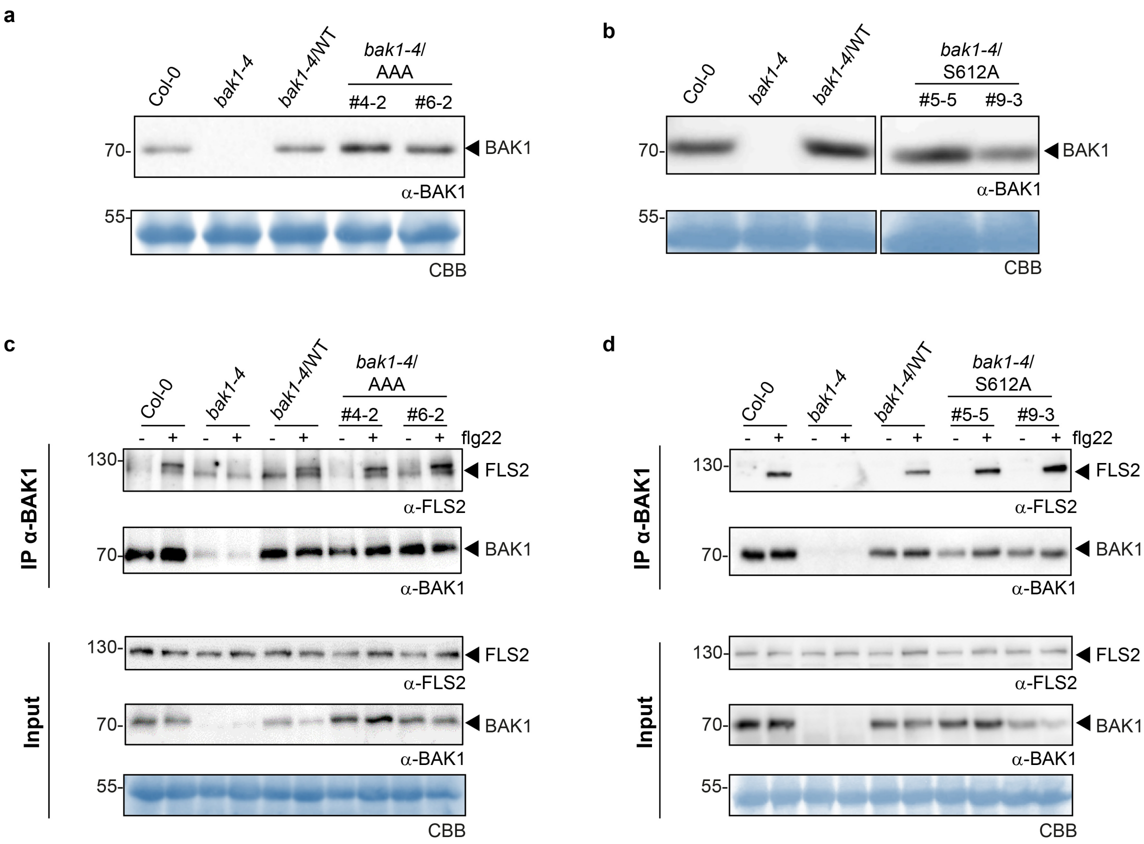
c

Position	Modified Peptide sequence	Sample Source	Peptide identification probability	Best Mascot Ion score
T446	LMDYKD(pT)HVTTAVR	FLS2-GFP	0.973	30.2
S602	D(pS)TSQIENEYPSGPR	FLS2-GFP and EFR-GFP IP	0.994	64.1
T603	DS(pT)SQIENEYPSGPR	FLS2-GFP and EFR-GFP IP	0.993	55.6
S604	DST(pS)QIENEYPSGPR	FLS2-GFP IP	0.994	36.5
S612	DSTSQIENEYP(pS)GPR	FLS2-GFP and EFR-GFP IP	0.997	57.4

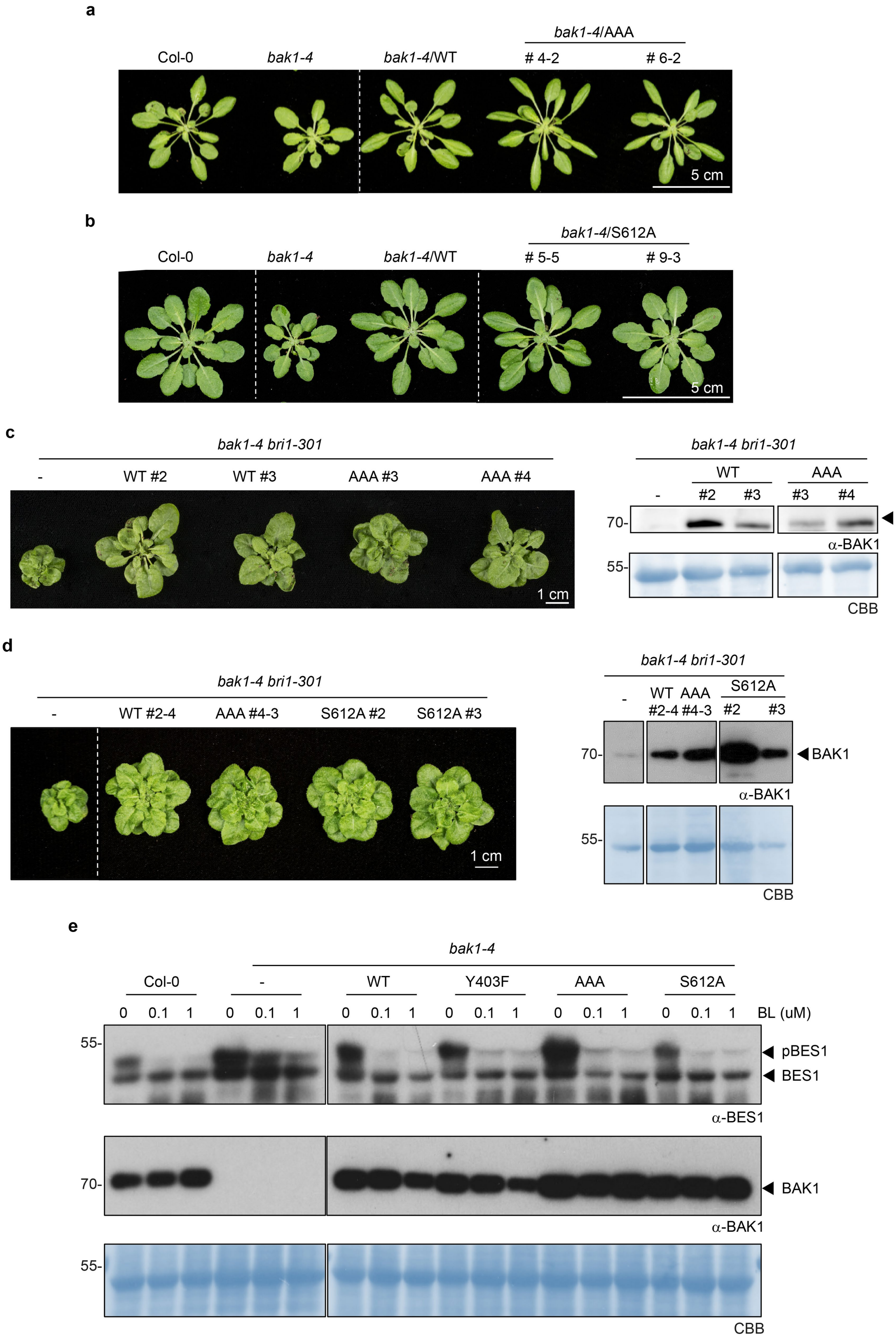
Extended Data Fig. 1 | Identification of BAK1 phosphosites. **a**, Representative CBB-stained SDS–PAGE gel showing proteins that are enriched upon GFP immunoprecipitation. **b**, Western blot of BAK1 co-immunoprecipitated with FLS2–GFP and EFR–GFP from **a** using anti-GFP and anti-BAK1 antibodies. For **a** and **b**, experiments were independently repeated three times. For gel and blot source data, see Supplementary Fig. 1. **c**, Summary of BAK1 in vivo phosphosites identified by FLS2–GFP and EFR–GFP co-immunoprecipitation followed by LC–MS/MS analysis. Spectra of identified sites are presented in Supplementary Fig. 2.



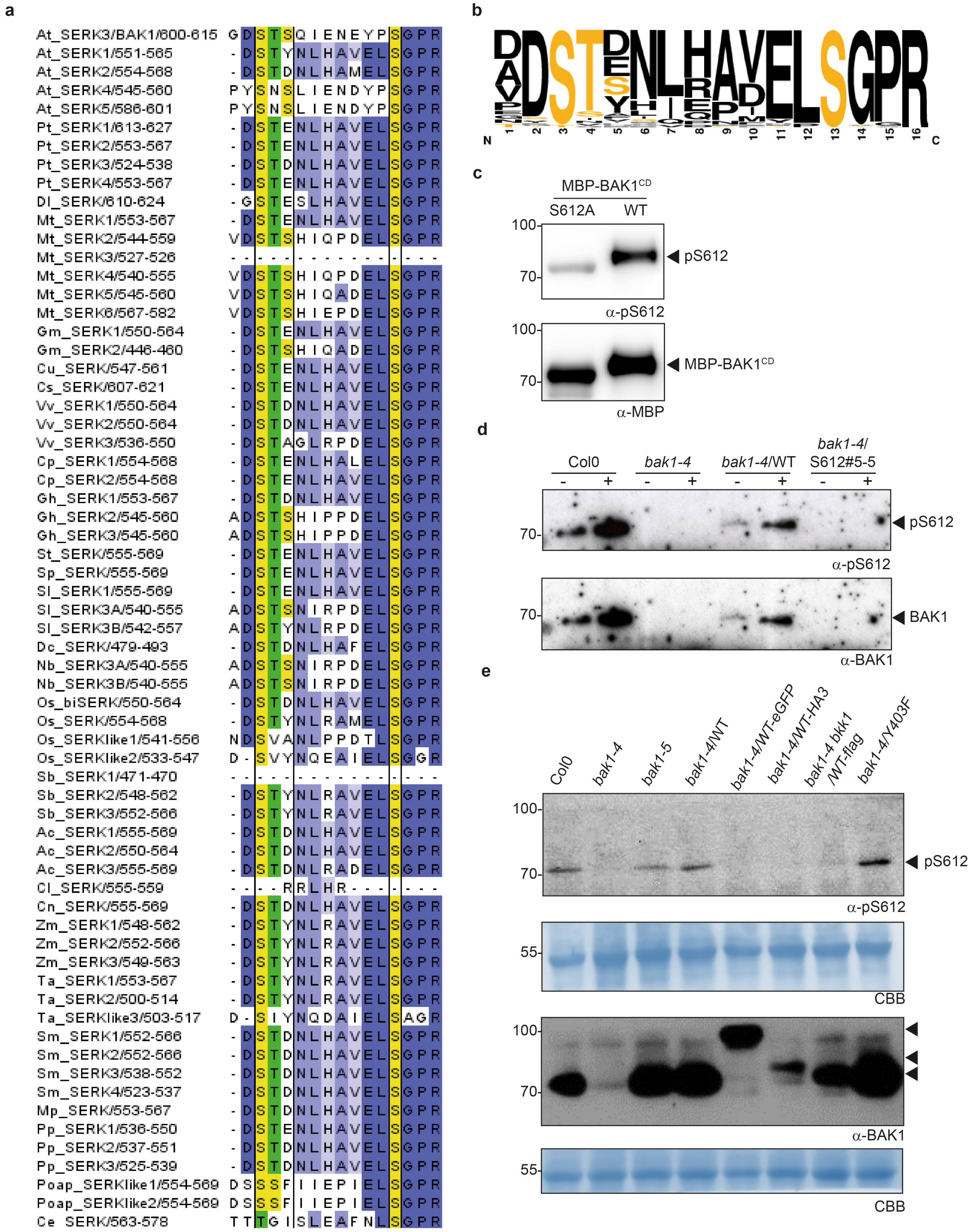
Extended Data Fig. 2 | Individual or double BAK1(S602A), BAK1(T603A) or BAK1(S604A) mutations do not affect elf18-induced ROS production. a, **e**, Total ROS production of *bak1-4* mesophyll protoplasts transiently expressing the indicated BAK1 mutants, after treatment with 100 nM elf18 for 60 min. $n = 12$ biological independent suspensions. **b**, **d**, **f**, Western blots with anti-BAK1 antibodies. For blot source data, see Supplementary Fig. 1. **c**, Total ROS production after treatment of leaf discs with 100 nM flg22 for 40 min. Individual data points are shown as circles. $n = 8$ biologically independent leaf discs. In **a**, **c** and **e**, measurements are plotted as box plots displaying the first and third quartiles, split by the median; whiskers extend to a maximum of 1.5× interquartile range beyond the box. Outliers are indicated as black dots. Statistical analysis was performed using a one-way ANOVA with Dunnett's post hoc test compared to wild-type BAK1. For **a**–**f**, experiments were repeated at least three times.


a
Col-0
bak1-4
bak1-4/WT
bak1-4/
AAA
#4-2
#6-2
70-
BAK1
α-BAK1
55-
CBB
b
Col-0
bak1-4
bak1-4/WT
bak1-4/
S612A
#5-5
#9-3
70-
BAK1
α-BAK1
55-
CBB
c
Col-0
bak1-4
bak1-4/WT
bak1-4/
AAA
#4-2
#6-2
- + - + - + - + - +
flg22
IP α-BAK1
130-
FLS2
α-FLS2
70-
BAK1
α-BAK1
Input
130-
FLS2
α-FLS2
70-
BAK1
α-BAK1
55-
CBB
d
Col-0
bak1-4
bak1-4/WT
bak1-4/
S612A
#5-5
#9-3
- + - + - + - + - +
flg22
IP α-BAK1
130-
FLS2
α-FLS2
70-
BAK1
α-BAK1
Input
130-
FLS2
α-FLS2
70-
BAK1
α-BAK1
55-
CBB

Extended Data Fig. 3 | BAK1(AAA) and BAK1(S612A) plants are not affected in flg22-induced FLS2–BAK1 complex formation. a, **b**, Western blots with anti-BAK1 antibodies. **c**, **d**, Co-immunoprecipitation between FLS2 and BAK1 using anti-FLS2 and anti-BAK1 antibodies. Two-week-old seedlings were treated with water (−) or 100 nM flg22 (+) for 10 min before protein extraction. Blots were stained with CBB for loading control. For **a**–**d**, experiments were repeated at least three times with similar results. For blot source data, see Supplementary Fig. 1.

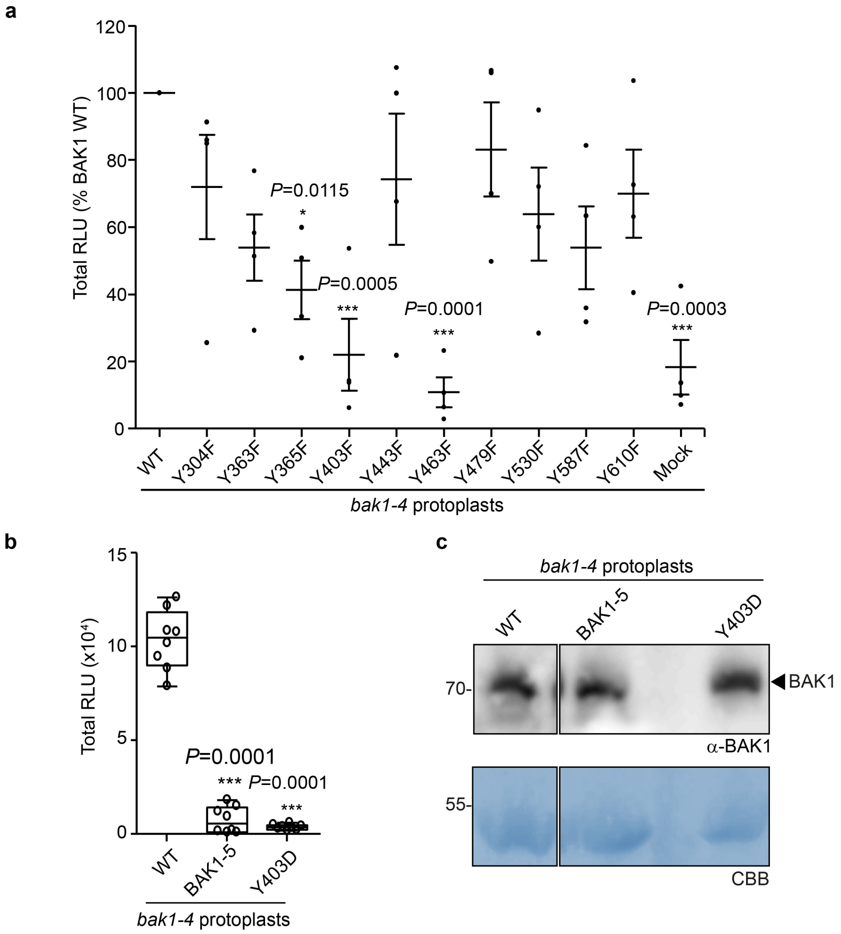

a
Col-0
bak1-4
bak1-4/WT
bak1-4/AAA
# 4-2
# 6-2
5 cm
b
Col-0
bak1-4
bak1-4/WT
bak1-4/S612A
# 5-5
# 9-3
5 cm
c
bak1-4 bri1-301
-
WT #2
WT #3
AAA #3
AAA #4
1 cm
bak1-4 bri1-301
WT
AAA
-
#2
#3
#3
#4
70-
BAK1
α-BAK1
55-
CBB
d
bak1-4 bri1-301
-
WT #2-4
AAA #4-3
S612A #2
S612A #3
1 cm
bak1-4 bri1-301
-
WT AAA
#2-4 #4-3
S612A
#2
#3
70-
BAK1
α-BAK1
55-
CBB
e
bak1-4
Col-0
-
WT
Y403F
AAA
S612A
0 0.1 1 0 0.1 1 0 0.1 1 0 0.1 1 0 0.1 1 0 0.1 1 BL (uM)
55-
pBES1
BES1
α-BES1
70-
BAK1
α-BAK1
55-
CBB

Extended Data Fig. 4 | Expression of the BAK1(AAA) and BAK1(S612A) variants restores the semi-dwarf rosette phenotype of *bak1-4* and *bak1-4 bri1-301*. a, b, Representative images of rosettes from four- to five-week-old plants. **c**, **d**, Left, representative images of rosettes from five- to seven-week-old plants. Right, western blots with anti-BAK1 antibodies. Blots were stained with CBB for loading control. Experiments were repeated at least twice with similar results. Control plants used in **b** have been previously pictured[30]. **e**, BES1 dephosphorylation 60 min after treatment of ten-day-old seedlings with the indicated concentrations of epi-BL (μM), as shown by western-blot analysis using anti-BES1 antibodies. Blots were probed with anti-BAK1 antibodies, and subsequently stained with CBB as loading controls. For **a**–**e**, experiments were repeated at least three times. For blot source data, see Supplementary Fig. 1.

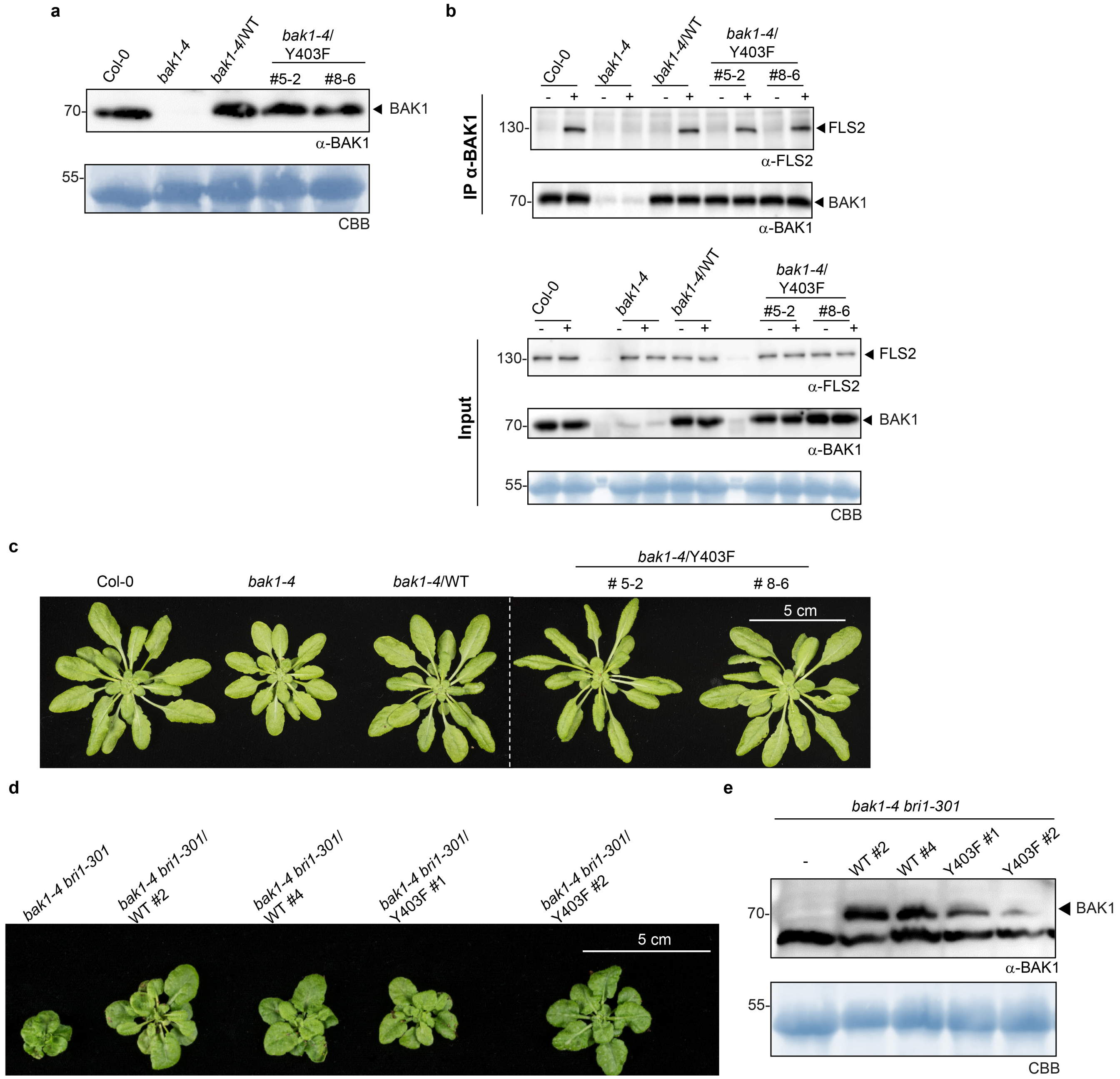


Extended Data Fig. 5 | See next page for caption.

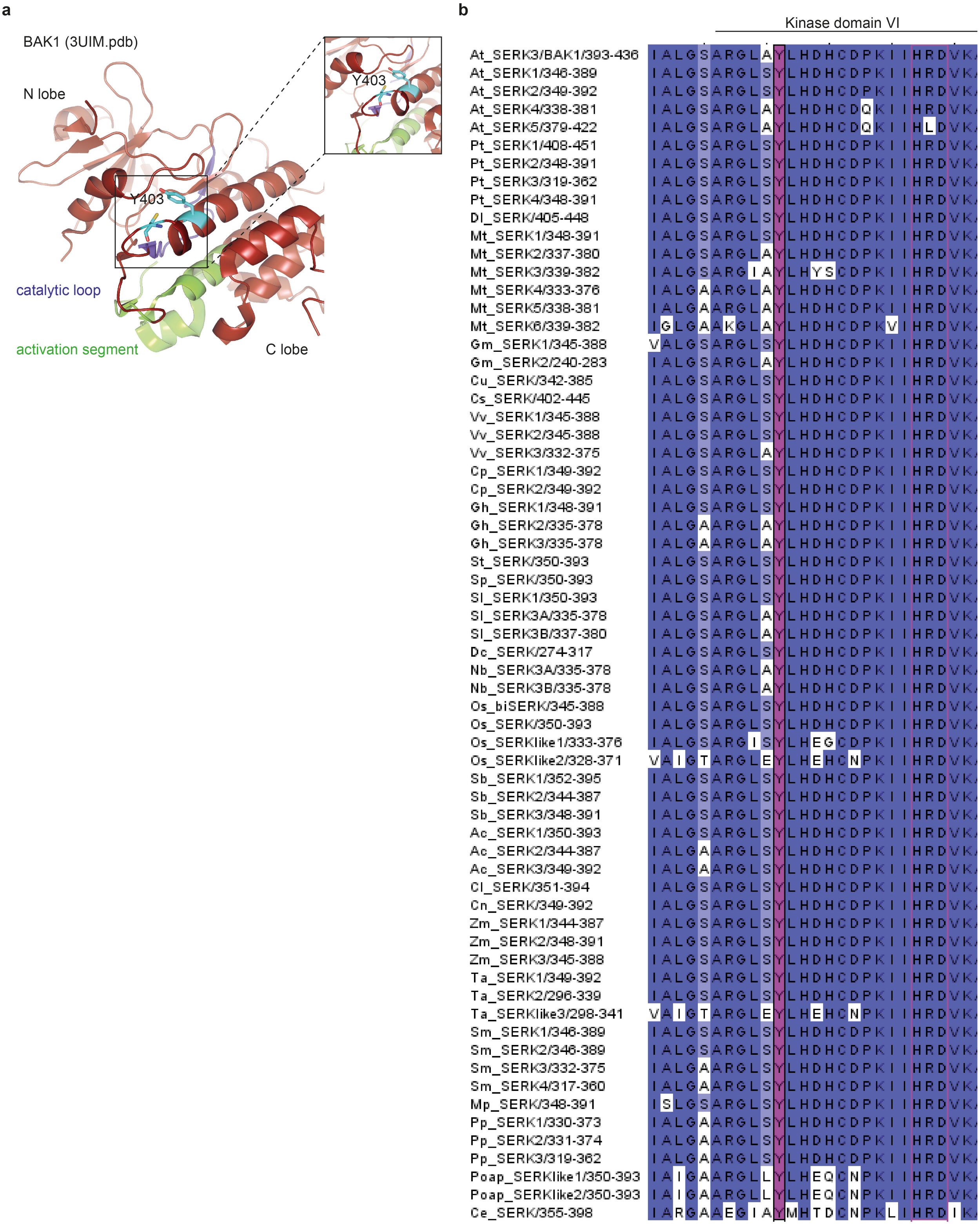
Extended Data Fig. 5 | Conservation of BAK1 C-terminal-tail phosphosites in SERK proteins across plant species. a, Clustal Omega multiple alignments were visualized using JalView v2.10.2b2. The alignment is coloured by percentage identity. Yellow, conservation of BAK1 S602, S604 and S612; green, conservation of BAK1 T603. Protein IDs used for the alignments are as follows: *Populus trichocarpa Pt*SERK1(B9MW41), *Pt*SERK2(B9IQM9), *Pt*SERK3(B9HFX1), *Pt*SERK4(B9H599), *Dimocarpus longan Dl*SERK(B5TTV0), *Medicago truncatula Mt*SERK1(Q8GRK2), *Mt*SERK2(E2IXG1), *Mt*SERK3(E2IXG8), *Mt*SERK4(E2IXG2), *Mt*SERK5(E2IXG3), *Mt*SERK6(E2IXG4), *Glycine max Gm*SERK1(C6ZGA8), *Gm*SERK2(C6FF61), *Citrus unshiu Cu*SERK (Q6BE26), *Citrus sinensis Cs*SERK(C3V9W0), *Vitis vinifera Vv*SERK1(D7TXV2), *Vv*SERK2(A5BIY4), *Vv*SERK3(D7STF6), *Cyclamen persicum Cp*SERK1(A7L5U3), *Cp*SERK2(E5D6S9), *Gossypium hirsutum Gh*SERK1(E5Q8K6), *Gh*SERK2(F5BZU9), *Gh*SERK3(F6MF11), *Solanum tuberosum St*SERK(A3R789), *Solanum peruvianum Sp*SERK(A6N8J2), *Solanum lycopersicum Sl*SERK1(G0XZA3), *Sl*SERK3A(G0XZA5), *Sl*SERK3B(G0XZA6), *Daucus carota Dc*SERK(O23921), *Arabidopsis thaliana At*SERK1(Q94AG2), *At*SERK2(Q9XIC7), *At*SERK3(Q94F62), *At*SERK4(Q9SKG5), *At*SERK5(Q8LPS5), *Nicotiana benthamiana Nb*SERK3A(E3VXE6), *Nb*SERK3B(E3VXE7), *Oryza sativa subsp. Indica Osbi*SERK(Q6S7F1), *Oryza sativa Os*SERK(Q5Y8C8), *Os*SERKlike1(Q67X31), *Os*SERKlike2(Q6K4T4), *Sorghum bicolor Sb*SERK1(C5YHV3), *Sb*SERK2(C5Y9S6), *Sb*SERK3(C5XVP5), *Ananas comosus Ac*SERK1(H6SU43), *Ac*SERK2(H6UP78), *Ac*SERK3(H6UP79), *Cyrtochilum loxense Cl*SERK(G2XLB1), *Cocos nucifera Cn*SERK(Q5S1N9), *Zea mays Zm*SERK1(Q93W70), *Zm*SERK2(Q94IJ5), *Zm*SERK3(B4G007), *Triticum aestivum Ta*SERK1(G4XGX1), *Ta*SERK2(G4XGX2), *Ta*SERKlike3(G4XGX3), *Selaginella moellendorffii Sm*SERK1(D8SBB8), *Sm*SERK2(D8S0N3), *Sm*SERK3(D8S4M4), *Sm*SERK4(D8R6C9), *Marchantia polymorpha Mp*SERK(A7VM18), *Physcomitrella patens Pp*SERK1(A9STU8), *Pp*SERK2(A9SMW5), *Pp*SERK3(A9RY79), *Poa pratensis Poap*SERKlike1(Q659J0), *Poap*SERKlike2(Q659J1), *Closterium ehrenbergii Ce*SERK(A7VM46). **b**, WebLogo representation of alignment in **a**. **c**, Detection of BAK1 S612 phosphorylation using anti-pS612 specific antibodies on affinity-purified recombinant BAK1 cytoplasmic domain (BAK1^{CD}) following an in vitro kinase assay with cold ATP. Membranes were immunoblotted with anti-pS612 and anti-MBP antibodies. **d**, BAK1 immunoprecipitation and detection of BAK1 S612 phosphorylation in vivo using anti-pS612 antibodies. Two-week-old seedlings were treated with water (−) or 100 nM flg22 (+) for 10 min. The same membrane was stripped and blotted again with anti-BAK1 antibodies for loading control. **e**, Western-blot analysis using anti-BAK1 and anti-pS612 antibodies after SDS–PAGE of crude protein extracts from two-week-old seedlings treated with 100 nM flg22 for 10 min. Blots were stained with CBB for loading control. For **a**–**e**, all experiments were repeated at least twice with similar results. For blot source data, see Supplementary Fig. 1.



Extended Data Fig. 6 | Mutational screen of tyrosine residues present in the cytoplasmic domain of BAK1. a, Total ROS production after treatment of *bak1-4* mesophyll protoplasts with 100 nM elf18 for 30 min. The RLU values are normalized relative to protoplasts expressing wild-type (WT) BAK1 and are expressed as aligned dot blots showing mean ± s.e.m. $n = 4$ independent biological experiments. **b**, Total ROS production of *bak1-4* mesophyll protoplasts that transiently express the indicated BAK1 mutants, after treatment with 100 nM elf18 for 60 min. $n = 8$ biologically independent suspensions. Measurements are plotted as box plots displaying the first and third quartiles, split by the median; whiskers extend to a maximum of 1.5× interquartile range beyond the box. Individual data points are shown as circles. **c**, Western blot with anti-BAK1 antibodies. For blot source data, see Supplementary Fig. 1. In **a** and **b**, statistical analysis was performed using a one-way ANOVA with Dunnett's post hoc test compared to the wild-type control. For **b** and **c**, experiments were repeated twice with similar results.

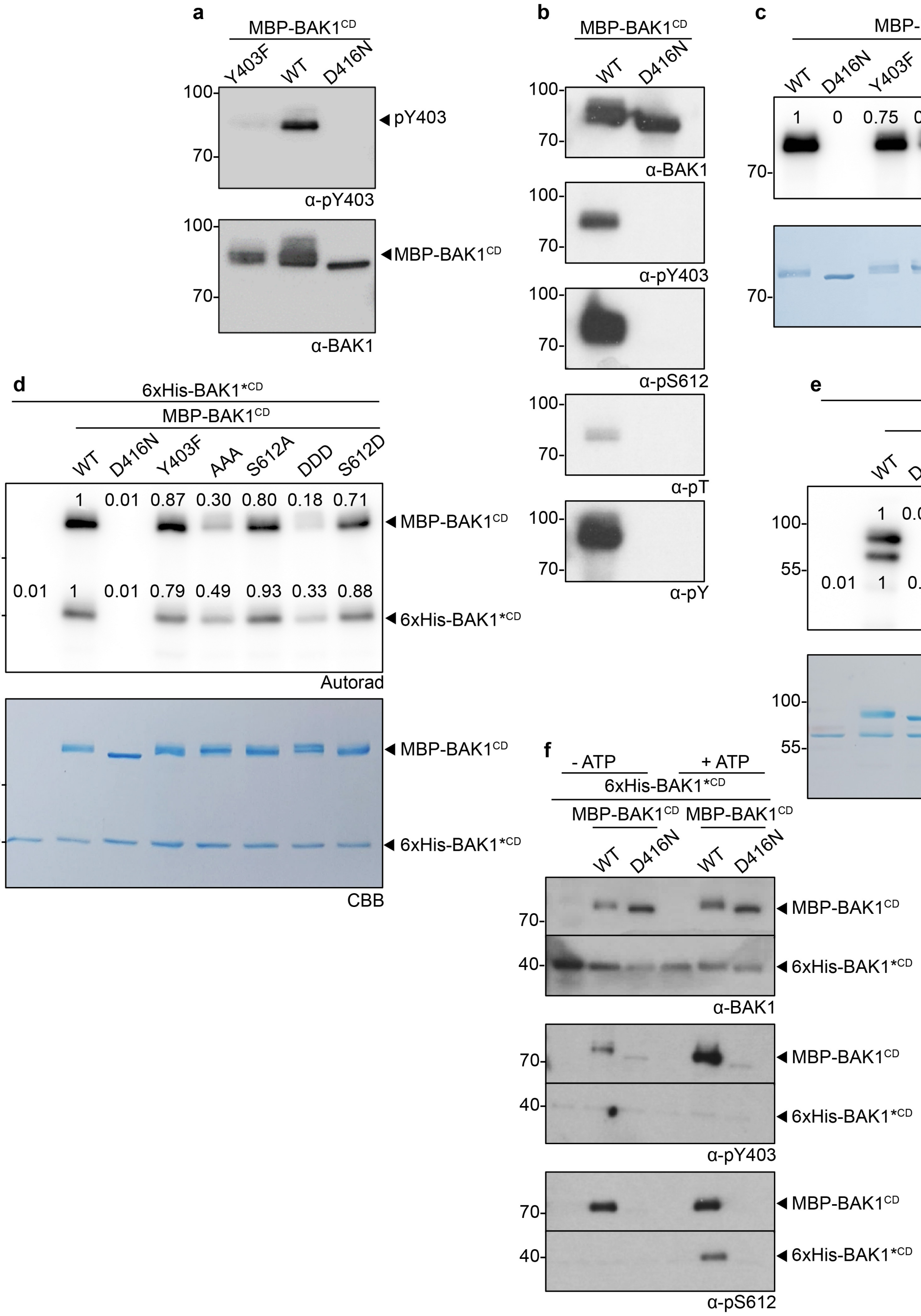

a
Col-0
bak1-4
bak1-4/WT
bak1-4/
Y403F
#5-2
#8-6
70-
BAK1
α-BAK1
55-
CBB
b
IP α-BAK1
Input
130-
FLS2
α-FLS2
70-
BAK1
α-BAK1
55-
CBB
c
Col-0
bak1-4
bak1-4/WT
bak1-4/Y403F
# 5-2
# 8-6
5 cm
d
bak1-4 bri1-301
bak1-4 bri1-301/
WT #2
bak1-4 bri1-301/
WT #4
bak1-4 bri1-301/
Y403F #1
bak1-4 bri1-301/
Y403F #2
5 cm
e
bak1-4 bri1-301
-
WT #2
WT #4
Y403F #1
Y403F #2
70-
BAK1
α-BAK1
55-
CBB

Extended Data Fig. 7 | BAK1 Y403 plants are not affected by flg22-induced FLS2–BAK1 complex formation and brassinosteroid-mediated rosette growth. a, Western blot with anti-BAK1 antibodies. Blots were stained with CBB for loading control. **b**, Co-immunoprecipitation experiments between FLS2 and BAK1 using anti-FLS2 and anti-BAK1 antibodies. Two-week-old seedlings were treated with water (−) or 100 nM flg22 (+) for 10 min before protein extraction. Blots were stained with CBB for loading control. **c**, Representative images of rosettes from four- to five-week-old plants. **d**, Representative images of rosettes from five-week-old plants. The same control plants used in **d** are also presented in Extended Data Fig. 4c. **e**, Western blot with anti-BAK1 antibodies. Blots were stained with CBB for loading control. For **a**–**e**, experiments were repeated three times with similar results. For blot source data, see Supplementary Fig. 1.

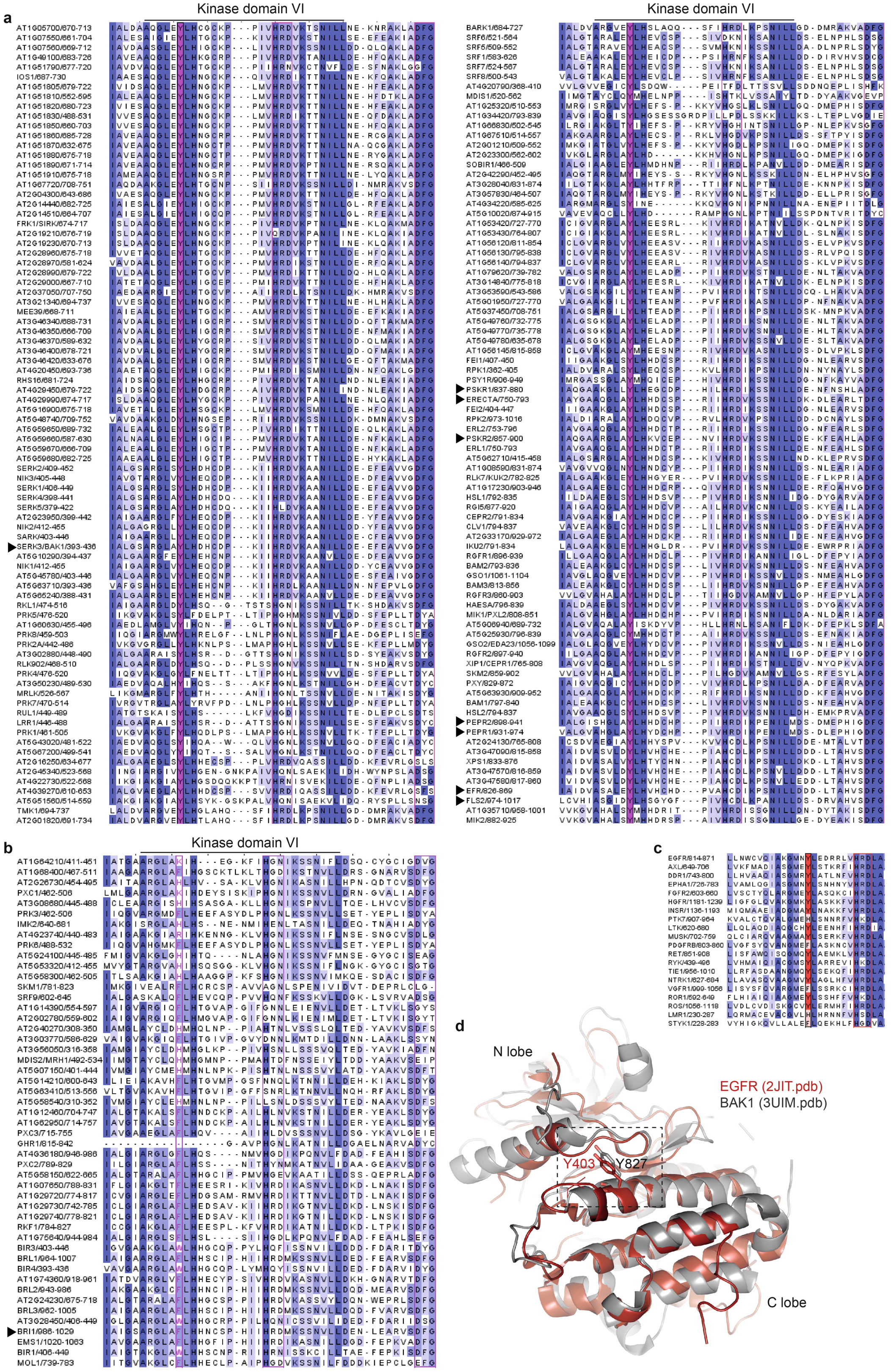

a
BAK1 (3UIM.pdb)
N lobe
Y403
Y403
catalytic loop
activation segment
C lobe
b
Kinase domain VI
At_SERK3/BAK1/393-436
At_SERK1/346-389
At_SERK2/349-392
At_SERK4/338-381
At_SERK5/379-422
Pt_SERK1/408-451
Pt_SERK2/348-391
Pt_SERK3/319-362
Pt_SERK4/348-391
Dl_SERK/405-448
Mt_SERK1/348-391
Mt_SERK2/337-380
Mt_SERK3/339-382
Mt_SERK4/333-376
Mt_SERK5/338-381
Mt_SERK6/339-382
Gm_SERK1/345-388
Gm_SERK2/240-283
Cu_SERK/342-385
Cs_SERK/402-445
Vv_SERK1/345-388
Vv_SERK2/345-388
Vv_SERK3/332-375
Cp_SERK1/349-392
Cp_SERK2/349-392
Gh_SERK1/348-391
Gh_SERK2/335-378
Gh_SERK3/335-378
St_SERK/350-393
Sp_SERK/350-393
Sl_SERK1/350-393
Sl_SERK3A/335-378
Sl_SERK3B/337-380
Dc_SERK/274-317
Nb_SERK3A/335-378
Nb_SERK3B/335-378
Os_biSERK/345-388
Os_SERK/350-393
Os_SERKlike1/333-376
Os_SERKlike2/328-371
Sb_SERK1/352-395
Sb_SERK2/344-387
Sb_SERK3/348-391
Ac_SERK1/350-393
Ac_SERK2/344-387
Ac_SERK3/349-392
Cl_SERK/351-394
Cn_SERK/349-392
Zm_SERK1/344-387
Zm_SERK2/348-391
Zm_SERK3/345-388
Ta_SERK1/349-392
Ta_SERK2/296-339
Ta_SERKlike3/298-341
Sm_SERK1/346-389
Sm_SERK2/346-389
Sm_SERK3/332-375
Sm_SERK4/317-360
Mp_SERK/348-391
Pp_SERK1/330-373
Pp_SERK2/331-374
Pp_SERK3/319-362
Poap_SERKlike1/350-393
Poap_SERKlike2/350-393
Ce_SERK/355-398

Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | The conserved Tyr-VIa residue BAK1 Y403 is in close proximity to the catalytic loop and to C408. a, In silico representation of the structure of BAK1^{CD} (PDB ID: 3UIM). The activation-segment region is shown in green and the catalytic loop in purple. **b**, Conservation of BAK1 Y403 in SERK proteins across plant species. Clustal Omega multiple alignments were visualized using JalView v2.10.2b2. The alignment is coloured by percentage identity. Magenta, conservation of BAK1 Y403. Protein IDs used for the alignments are as follows: PpSERK1(B9MW41), PpSERK2(B9IQM9), PpSERK3(B9HFX1), PpSERK4(B9H599), DlSERK(B5TTV0), MtSERK1(Q8GRK2), MtSERK2(E2IXG1), MtSERK3(E2IXG8), MtSERK4(E2IXG2), MtSERK5(E2IXG3), MtSERK6(E2IXG4), GmSERK1(C6ZGA8), GmSERK2(C6FF61), CuSERK (Q6BE26), CsSERK(C3V9W0), VvSERK1(D7TXV2), VvSERK2(A5BIY4), VvSERK3(D7STF6), CpSERK1(A7L5U3), CpSERK2(E5D6S9), GhSERK1(E5Q8K6), GhSERK2(F5BZU9), GhSERK3(F6MF11), StSERK(A3R789), SpSERK(A6N8J2), SlSERK1(G0XZA3), SlSERK3A(G0XZA5), SlSERK3B(G0XZA6), DcSERK(O23921), AtSERK1(Q94AG2), AtSERK2(Q9XIC7), AtSERK3(Q94F62), AtSERK4(Q9SKG5), AtSERK5(Q8LPS5), NbSERK3A(E3VXE6), NbSERK3B(E3VXE7), OsbiSERK(Q6S7F1), OsSERK(Q5Y8C8), OsSERKlike1(Q67X31), OsSERKlike2(Q6K4T4), SbSERK1(C5YHV3), SbSERK2(C5Y9S6), SbSERK3(C5XVP5), AcSERK1(H6SU43), AcSERK2(H6UP78), AcSERK3(H6UP79), ClSERK(G2XLB1), CnSERK(Q5S1N9), ZmSERK1(Q93W70), ZmSERK2(Q94IJ5), ZmSERK3(B4G007), TaSERK1(G4XGX1), TaSERK2(G4XGX2), TaSERKlike3(G4XGX3), SmSERK1(D8SBB8), SmSERK2(D8S0N3), SmSERK3(D8S4M4), SmSERK4(D8R6C9), MpSERK(A7VM18), PpSERK1(A9STU8), PpSERK2(A9SMW5), PpSERK3(A9RY79), PoapSERKlike1(Q659J0), PoapSERKlike2(Q659J1), CeSERK(A7VM46).



Extended Data Fig. 9 | Analysis of phosphosite-specific kinase activities of BAK1. a, Specific detection of BAK1 Y403 phosphorylation on affinity-purified recombinant BAK1^{CD} wild-type but not Y403F or kinase-dead (D416N) proteins, using anti-pY403 antibodies and following an in vitro kinase assay with cold ATP. Blots were probed with anti-BAK1 antibodies for loading control. **b**, Detection of BAK1 Y403, S612, threonine and tyrosine phosphorylation using phosphorylation-specific antibodies on affinity-purified recombinant BAK1^{CD} following an in vitro kinase assay with cold ATP. Membranes were immunoblotted with BAK1 antibodies for loading control. **c**, [^{32}P]γ-ATP kinase assays showing autophosphorylation of the indicated MBP-BAK1^{CD} phosphosite mutants. **d**, *trans*-autophosphorylation of the kinase-dead BAK1(D416N) (6×His-BAK1*^{CD}) by the indicated BAK1 phosphosite mutants (MBP–BAK1^{CD}). **e**, *Trans*-phosphorylation activity of the indicated affinity purified recombinant BAK1^{CD} mutants against a kinase-dead (K105E) BIK1 substrate (GST–BIK1*). **f**, *Trans*-autophosphorylation of kinase-dead (D416N) BAK1 substrate (6×His-BAK1*^{CD}) at S612 and Y403 by wild-type BAK1 (MBP–BAK1^{CD}). BAK1 Y403 or S612 phosphorylation was detected using phosphorylation-specific antibodies following an in vitro incubation of the purified proteins in the presence or in the absence of cold ATP. Membranes were immunoblotted with BAK1 antibodies for loading control. In **c**–**e**, numbers indicate autoradiograph band intensity relative to the wild type. Protein loading control was determined with CBB staining. For **a**–**f**, all experiments were repeated independently at least three times. For blot source data, see Supplementary Fig. 1.


a
Kinase domain VI
Kinase domain VI
SERK3/BAK1/393-436
PSKR1/837-880
ERECTA/750-793
PSKR2/857-900
PEPR2/898-941
PEPR1/931-974
EFR/826-869
FLS2/974-1017
b
Kinase domain VI
BRI1/986-1029
c
d
N lobe
EGFR (2JIT.pdb)
BAK1 (3UIM.pdb)
Y403
Y827
C lobe

Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | The conservation of Tyr-VIa in the *A. thaliana* LRR-RK cytoplasmic domain and in representative members of the 20 groups of animal-receptor tyrosine kinases. a, **b**, Clustal Omega multiple alignments of *A. thaliana* LRR-RK cytoplasmic domain, visualized using JalView v2.10.2b2, illustrate the conservation (**a**) and non-conservation (**b**) of Tyr-VIa. The alignment is coloured by percentage identity. Magenta, Tyr-VIa. Protein IDs of the sequences used for the alignment are in Supplementary Table 1. Arrows indicate the LRR-RKs reported in this study. **c**, The conservation of Tyr-VIa in representative members of the 20 groups of animal receptor tyrosine kinases. Clustal Omega multiple alignments were visualized using JalView v2.10.2b2. The alignment is coloured by percentage identity. Red, position analogous to BAK1 Y403. Protein IDs used for the alignments: EGFR (P00533), AXL (P30530), DDR15 (Q08345), EphA1 (P21709), FGFR2 (P21802), HGFR (P08581), INSR (P06213), PTK7 (Q13308), LTK (P29376), MUSK (O15146), PGFRB (P09619), RET (P07949), RYK (P34925), TIE1 (P35590), NTRK1 (P04629), VGFR1 (P17948), ROR1 (Q01973), ROS (P08922), LMR1 (Q6ZMQ8), STYK1 (Q6J9G0). **d**, In silico alignments of the structures of BAK1^{CD} and EGFR^{CD}. A selected overlapping region of the BAK1 (PDB ID: 3UIM) and EGFR (PDB ID: 2JIT) cytoplasmic domain structures is presented, highlighting BAK1 Y403 and EGFR Y827.

LETTER

https://doi.org/10.1038/s41586-018-0484-5

An orthogonal proteomic survey uncovers novel Zika virus host factors

Pietro Scaturro[1,2]*, Alexey Stukalov[1,2], Darya A. Haas[1], Mirko Cortese[3], Kalina Draganova[4,5], Anna Płaszczyca[3], Ralf Bartenschlager[3,6], Magdalena Götz[4,5,7] & Andreas Pichlmair[1,2,8]*

Zika virus (ZIKV) has recently emerged as a global health concern owing to its widespread diffusion and its association with severe neurological symptoms and microcephaly in newborns[1]. However, the molecular mechanisms that are responsible for the pathogenicity of ZIKV remain largely unknown. Here we use human neural progenitor cells and the neuronal cell line SK-N-BE2 in an integrated proteomics approach to characterize the cellular responses to viral infection at the proteome and phosphoproteome level, and use affinity proteomics to identify cellular targets of ZIKV proteins. Using this approach, we identify 386 ZIKV-interacting proteins, ZIKV-specific and pan-flaviviral activities as well as host factors with known functions in neuronal development, retinal defects and infertility. Moreover, our analysis identified 1,216 phosphorylation sites that are specifically up- or downregulated after ZIKV infection, indicating profound modulation of fundamental signalling pathways such as AKT, MAPK–ERK and ATM–ATR and thereby providing mechanistic insights into the proliferation arrest elicited by ZIKV infection. Functionally, our integrative study identifies ZIKV host-dependency factors and provides a comprehensive framework for a system-level understanding of ZIKV-induced perturbations at the levels of proteins and cellular pathways.

ZIKV, a flavivirus which is related to dengue virus, West Nile virus and hepatitis C virus (HCV), has a single-stranded RNA genome of positive polarity, encoding a polyprotein that is co- and post-translationally processed into three structural (capsid (C), precursor membrane (prM) and envelope (E)) and seven non-structural proteins (NS1, NS2A, NS2B, NS3, NS4A, NS4B and NS5)[2]. To comprehensively understand how ZIKV affects neuronal cells, we performed an unbiased proteomic survey to identify cellular proteins and associated complexes that interact with each of the ten ZIKV proteins expressed in human SK-N-BE2 neuroblastoma cells (Extended Data Fig. 1a). Except for NS2A, all ZIKV proteins were correctly expressed and processed (Extended Data Fig. 1b, c). Affinity purification coupled with liquid chromatography and tandem mass spectrometry (AP–LC–MS/MS), followed by Bayesian statistical modelling, identified 386 proteins specifically associating with ZIKV proteins, resulting in 484 high-confident interactions (Fig. 1, Extended Data Fig. 2 and Supplementary Table 1). We identified previously reported proteins that showed bona fide interactions with flavivirus proteins, including subunits of the ATPase (ATP1A1, ATP1A2, ATP1A3, ATP1B and ATP6V1H), voltage-dependent anion-selective channel proteins (VDAC1, VDAC2 and VDAC3) as well as components of the cytochrome *c* oxidase complex (COX15, MT-CO2 and NDUFA4), confirming these processes as important targets of diverse flaviviruses[3] (Fig. 1 and Extended Data Fig. 2). Notably, this analysis uncovered proteins linked to neurological diseases or development, particularly among the proteins that specifically interacted with capsid and NS4B (Fig. 1). For instance, the capsid-interacting proteins included LARP7 (involved in telomere stability and mutated in Alazami syndrome)[4], LYAR (important for maintenance of embryonic stem cell identity)[5] and NGDN (a neuronal development factor)[6] (Extended Data Fig. 3a). NS4B has recently been implicated in the inhibition of neuronal development[7]. Compared to HCV-NS4B, ZIKV-NS4B showed specific enrichment in cellular proteins associated with the whole spectrum of ZIKV-associated pathogenesis, ranging from neurodegenerative disorders and retinal degeneration (CLN6 and BSG)[8,9] to regulators of neuronal differentiation (CEND1 and RBFOX2)[10,11] and axonal dysfunction (CHP1 and TMEM41b)[12] (Fig. 1 and Extended Data Figs. 2–4a). Co-immunoprecipitation followed by western blotting of transduced or ZIKV-infected cells verified that these proteins specifically associate with arthropod-borne NS4Bs (such as those of dengue virus and ZIKV) and not NS4B of other *Flaviviridae*, such as HCV; with a subset of proteins specifically binding to ZIKV-NS4B only (TMEM41b, CEND1 and CLN6; Extended Data Fig. 4b, c). ZIKV-NS4B precipitated particularly well with ceroid-lipofuscinosis neuronal protein 6 (CLN6), which is associated with a lysosomal storage disease that causes neurodegenerative late-infantile disorders as well as retinal defects[8]. In human neural progenitor cells (hNPCs), CLN6 redistributed to sites enriched in NS4B (Extended Data Fig. 3d). AP–LC–MS/MS experiments using CLN6 as bait identified common binding partners between NS4B and CLN6 (Extended Data Fig. 3e). Furthermore, CLN6 associated specifically with mTOR and TELO2, important regulators of signalling pathways that are disrupted by ZIKV[7,13]. Taken together, the ZIKV interactome revealed a number of new Flavivirus host-binding partners, including cellular proteins and signalling pathway components involved in neuronal development.

In patients, during early stages of development, ZIKV predominantly infects NPCs, causing microcephaly and other neurodevelopmental injuries[14]. To gain insights into ZIKV-induced changes, we used a human induced pluripotent stem cell-derived neuronal differentiation model[15]. Analysis of the global proteomic changes that occurred during differentiation of hNPCs into neurons revealed significant upregulation of neuronal differentiation markers such as βIII-tubulin (TUBB3), microtubule-associated proteins 2 and 6 (MAP2 and MAP6), neuronal cell adhesion molecule 1 (NCAM1), doublecortin (DCX) and ELAV-like RNA binding protein 3 (ELAVL3) (Fig. 2a, Extended Data Fig. 5 and Supplementary Table 2). Notably, ectopic expression of ZIKV-NS4B during differentiation specifically downregulated the expression of a subset of proteins involved in neuronal differentiation (for example, MAP2, MAP6, DPYSL3, DPYSL5 and CNTN2) as well as proteins associated with neurological diseases (for example, DOK3 and SUMO2), suggesting disruption of specific developmental programs (Fig. 2b, c, Extended Data Fig. 5c–e and Supplementary Tables 2, 3). Similar effects could be seen in proteomic analysis of hNPCs infected with ZIKV in the presence or absence of differentiation stimuli (Extended Data Fig. 6a). ZIKV infection led to a robust upregulation of type-I interferon-stimulated genes (for example, STAT1, MX1, OAS3 and IFIT1)

---

[1]Max-Planck Institute of Biochemistry, Innate Immunity Laboratory, Martinsried, Germany. [2]Technical University of Munich, School of Medicine, Institute of Virology, Munich, Germany. [3]Department of Infectious Diseases, Molecular Virology, University of Heidelberg, Heidelberg, Germany. [4]Institute of Stem Cell Research, Helmholtz Center Munich, Neuherberg, Germany. [5]Physiological Genomics, Biomedical Center, Ludwig-Maximilians-Universitaet, Munich, Germany. [6]German Center for Infection Research (DZIF), Heidelberg partner site, Heidelberg, Germany. [7]Synergy, Excellence Cluster for Systems Neurology, Biomedical Center, Ludwig-Maximilians-Universitaet, Munich, Germany. [8]German Center for Infection Research (DZIF), Munich partner site, Munich, Germany. *e-mail: pietro.scaturro@tum.de; andreas.pichlmair@tum.de

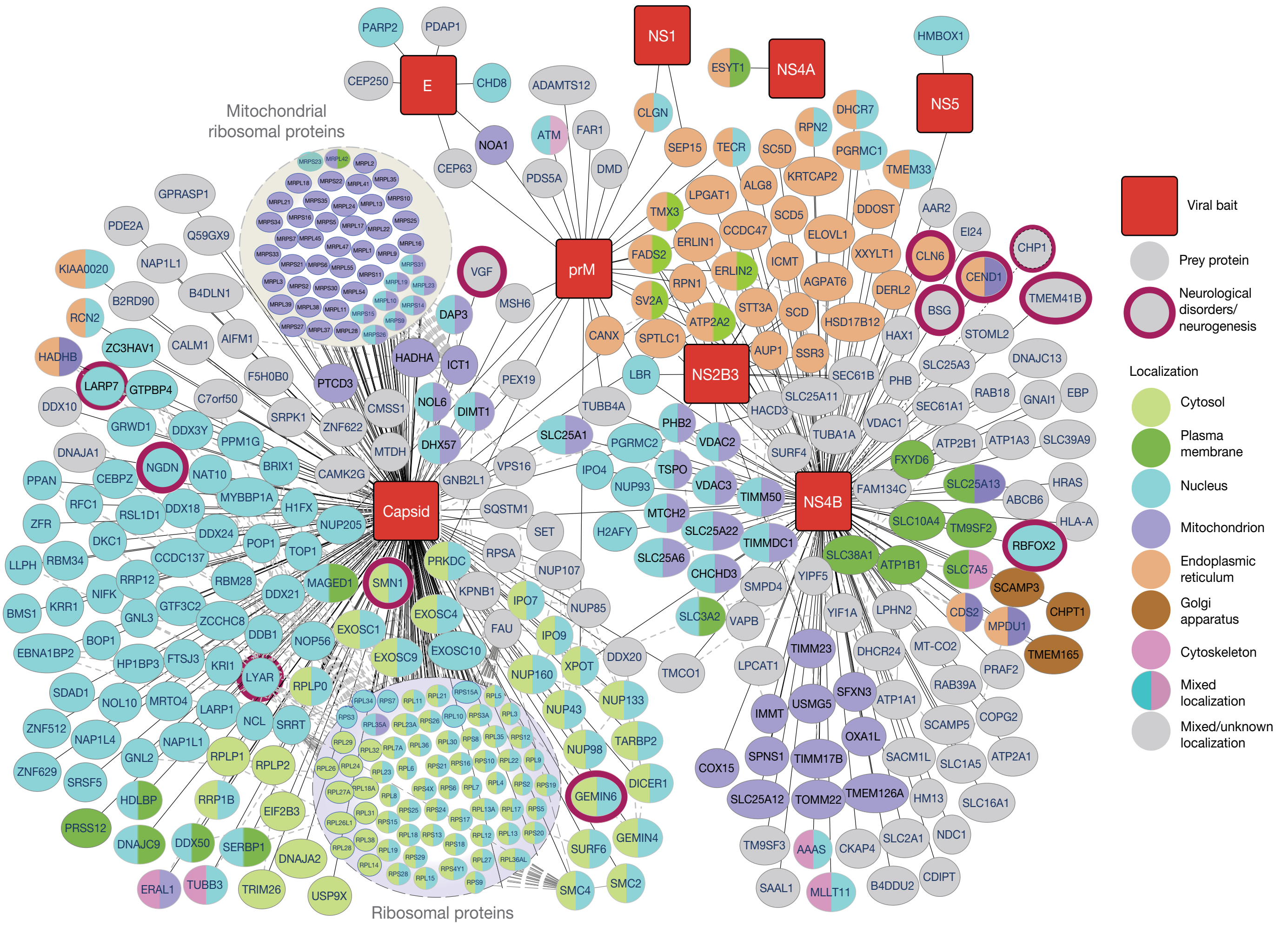

PARP2
PDAP1
E
CEP250
CHD8
ADAMTS12
NS1
ESYT1
NS4A
HMBOX1
NS5
Mitochondrial ribosomal proteins
CLGN
ATM
FAR1
NOA1
CEP63
PDS5A
DMD
SEP15
TECR
SC5D
RPN2
DHCR7
PGRMC1
TMEM33
KRTCAP2
ALG8
LPGAT1
TMX3
DDOST
AAR2
GPRASP1
PDE2A
Q59GX9
KIAA0020
NAP1L1
B2RD90
B4DLN1
RCN2
VGF
prM
FADS2
ERLIN1
CCDC47
SCD5
ELOVL1
ICMT
XXYLT1
CLN6
EI24
CHP1
CEND1
TMEM41B
SV2A
RPN1
ERLIN2
AGPAT6
DERL2
BSG
MSH6
DAP3
CANX
STT3A
SCD
HSD17B12
HAX1
STOML2
SPTLC1
ATP2A2
AUP1
SSR3
HADHB
ZC3HAV1
CALM1
AIFM1
HADHA
ICT1
LARP7
GTPBP4
F5H0B0
PTCD3
PEX19
LBR
NS2B3
SEC61B
PHB
SLC25A3
DNAJC13
DDX10
C7orf50
SRPK1
CMSS1
NOL6
DIMT1
TUBB4A
HACD3
SLC25A11
SEC61A1
RAB18
GNAI1
EBP
GRWD1
DDX3Y
PPM1G
ZNF622
DHX57
SLC25A1
PHB2
TUBA1A
VDAC1
ATP2B1
ATP1A3
SLC39A9
DNAJA1
NGDN
NAT10
BRIX1
MTDH
CAMK2G
VPS16
PGRMC2
VDAC2
SURF4
FXYD6
PPAN
CEBPZ
GNB2L1
IPO4
TSPO
NUP93
VDAC3
TIMM50
NS4B
FAM134C
SLC25A13
HRAS
RFC1
MYBBP1A
H1FX
Capsid
SQSTM1
ABCB6
ZFR
RSL1D1
DDX18
NUP205
MTCH2
SLC10A4
TM9SF2
HLA-A
DKC1
DDX24
POP1
SET
H2AFY
SLC25A22
TIMMDC1
RBFOX2
LLPH
RBM34
CCDC137
TOP1
RPSA
SLC25A6
SLC38A1
ATP1B1
SLC7A5
SCAMP3
RRP12
RBM28
MAGED1
SMN1
PRKDC
NUP107
CHCHD3
YIPF5
SMPD4
NIFK
DDX21
KPNB1
CDS2
CHPT1
BMS1
KRR1
GNL3
GTF3C2
ZCCHC8
EXOSC4
IPO7
NUP85
SLC3A2
VAPB
YIF1A
LPHN2
MPDU1
EBNA1BP2
BOP1
DDB1
NOP56
EXOSC1
FAU
IPO9
TMEM165
HP1BP3
FTSJ3
KRI1
EXOSC9
EXOSC10
XPOT
DDX20
TMCO1
LPCAT1
TIMM23
DHCR24
MT-CO2
PRAF2
SDAD1
NOL10
MRTO4
LYAR
RPLP0
NUP160
SFXN3
RAB39A
ZNF512
LARP1
NCL
SRRT
NUP43
NUP133
IMMT
USMG5
ATP1A1
NAP1L4
NAP1L1
GNL2
RPLP1
RPLP2
NUP98
TARBP2
OXA1L
SCAMP5
COPG2
ZNF629
SRSF5
COX15
SPNS1
TIMM17B
SACM1L
SLC1A5
ATP2A1
HDLBP
RRP1B
EIF2B3
GEMIN6
DICER1
SLC25A12
TOMM22
TMEM126A
HM13
SLC16A1
PRSS12
DNAJC9
DDX50
SERBP1
DNAJA2
SURF6
GEMIN4
TM9SF3
AAAS
CKAP4
SLC2A1
NDC1
ERAL1
TUBB3
TRIM26
USP9X
SMC4
SMC2
SAAL1
MLLT11
B4DDU2
CDIPT
Ribosomal proteins
Viral bait
Prey protein
Neurological disorders/ neurogenesis
Localization
Cytosol
Plasma membrane
Nucleus
Mitochondrion
Endoplasmic reticulum
Golgi apparatus
Cytoskeleton
Mixed localization
Mixed/unknown localization

Fig. 1 | A ZIKV proteins–host proteins interaction network in neuronal cells. Network representation of the high-confidence ZIKV–host interactome in SK-N-BE2 cells. There are eight ZIKV bait proteins (red squares) and 386 interacting host proteins (ellipses). $n = 4$ independent experiments, Bayesian statistical modelling. All baits: $\log_2$(fold change) ≥ 2.5; unadjusted one-sided $P \leq 0.05$; capsid: $\log_2$(fold change) ≥ 2.5; unadjusted one-sided $P \leq 5 \times 10^{-5}$. Interactions between ZIKV and host proteins are indicated by black lines. Published physical associations of the host proteins are indicated by dotted grey lines. Subcellular localization (colour of nodes) of host proteins is taken from Gene Ontology. Proteins with known functions in neurogenesis or neurological diseases are circled in pink.

alongside specific downregulation of neuronal factors under differentiation conditions (Extended Data Fig. 6b and Supplementary Table 4). These experiments unveil the distinct effects of ZIKV and ZIKV-NS4B on the proteome of differentiating hNPCs, revealing specific targets that are differentially regulated during neurogenesis.

Post-translational modifications of proteins, especially phosphorylation, have crucial roles in controlling the complex reorganization of signal-transduction pathways upon virus infections[16,17]. We therefore used LC–MS/MS-based time-resolved phosphoproteomics[18] to investigate cellular pathway deregulation in ZIKV-infected SK-N-BE2 cells. This analysis identified 14,222 unique phosphorylation sites (localization probability ≥ 0.75; Extended Data Fig. 7a–d), 1,216 of which were modulated upon ZIKV infection (Fig. 2d and Supplementary Table 5). Analysis of enriched biological functions showed that pathways involved in cellular assembly and organization, cell-cycle regulation, nervous system development and neurological diseases were significantly modulated by ZIKV infection (Extended Data Fig. 8a and Supplementary Table 6). Notably, this approach allowed us to map specific phosphorylation events along entire cellular pathways, highlighting modulation of key processes such as ATM, AKT–mTOR and ERK–MAPK signalling cascades (Extended Data Figs. 7e, 8b and Supplementary Tables 7, 8). We identified strong downregulation of the AKT–mTOR signalling pathway, as indicated by dephosphorylation of known AKT1 substrates (for example, DNMT1, TBC1D4 and LARP6), mTOR targets (for example, ANKRD17, LARP1, PATL1 and EEF2K), the central kinase S6K and its main effector protein S6. Consistently, we found increased phosphorylation of proteins, such as the negative regulator of autophagy DAP, at residues known to inhibit its suppressor function and consistent with the notion that autophagy is induced after ZIKV infection[7,19] (Fig. 2e and Extended Data Fig. 7f). Notably, ZIKV infection additionally downregulated the MAPK–ERK-signalling pathway, since A-RAF and ERK2 as well as several substrates of the downstream ERK1/2 map kinases (EIF4EBP1, BAZ1B and TWIST) were significantly dephosphorylated upon ZIKV infection (Fig. 2e, Extended Data Fig. 8c, d and Supplementary Tables 5, 8). These data enable us to quantitatively map individual phosphorylated residues onto signalling pathways on a global scale, identifying complex interactions between different cellular pathways (Extended Data Fig. 7e). For instance, both ERK1/2 and mTOR pathways converge on S6K and modulate its phosphorylation status[20] (Fig. 2e), suggesting multiple possible mechanisms that could contribute to the observed ZIKV-induced upregulation of autophagy, decrease in hNPC proliferation and impairment of neurogenesis. Notably, upregulation of the ATM (ataxia-telangiectasia mutated)/ATR (ATM- and Rad3-related) DNA-damage pathway, as inferred from significantly increased phosphorylation of several substrates of ATR (NPM1, CGGBP and GINS2), DNAPK (XRCC4) and downstream effector proteins (H2AFX), as well as proteins involved in cell-cycle regulation and DNA-damage checkpoints (TOP2B, CDK1)

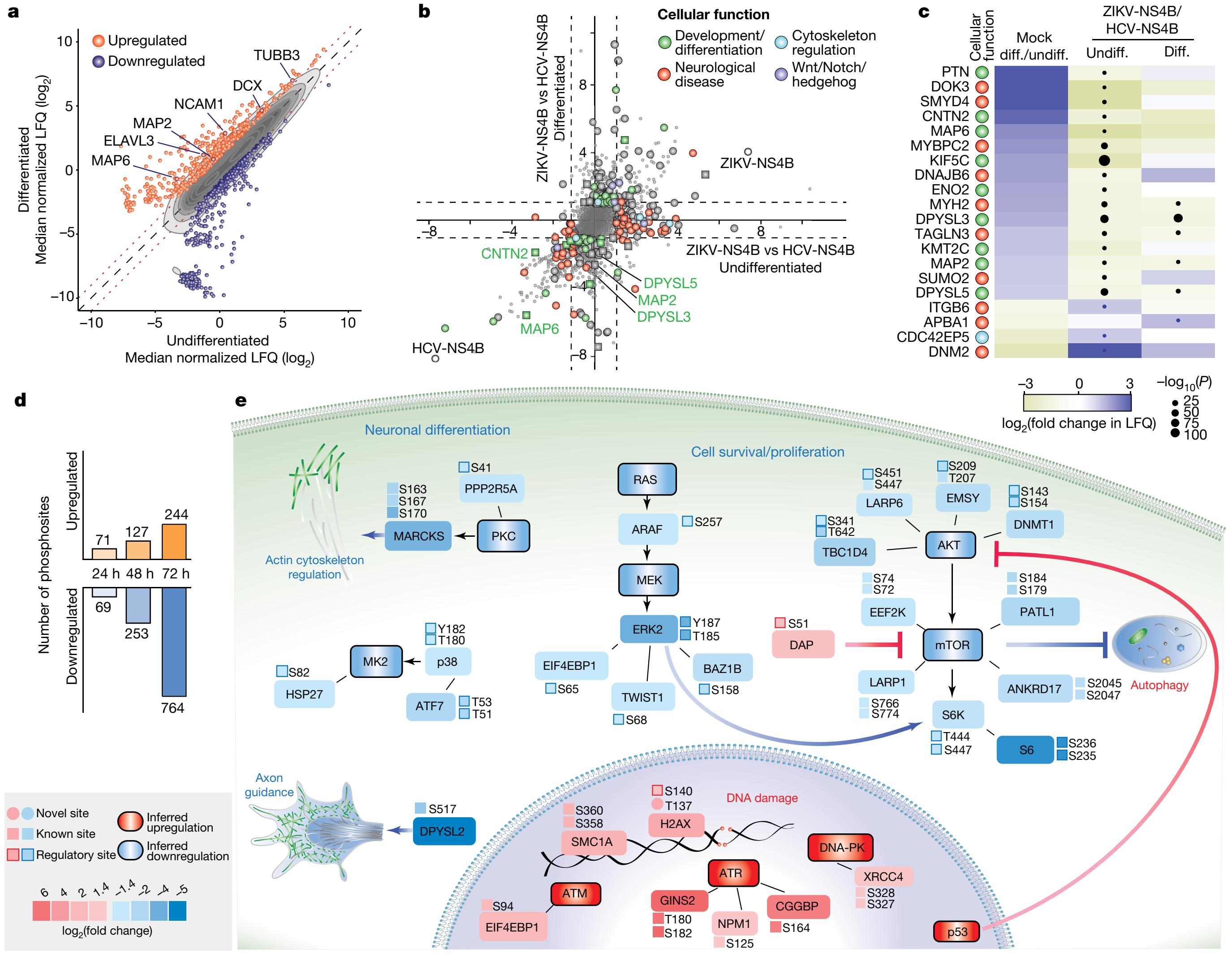


Fig. 2 | Impact of ZIKV and NS4B on hNPCs proteome upon differentiation and global phosphoproteomic profiling of ZIKV-infected cells. a, Comparative proteomic analysis of hNPCs, undifferentiated or differentiated for five days in the presence of ROCK inhibitor. LFQ, label-free quantification. Significantly up- or downregulated proteins are shown in orange and blue, respectively. $n = 4$ independent experiments, Bayesian statistical modelling. Unadjusted two-sided $P \leq 0.01$; $|\log_2(\text{fold change})| \geq 1$. Known markers of neuronal differentiation are highlighted. **b**, Impact of ZIKV-NS4B on differentiated and undifferentiated hNPCs. The scatter plot displays ZIKV-NS4B specific changes (compared to HCV-NS4B). Proteins that were specifically modulated by ZIKV-NS4B are shown as large circles or squares. Colour codes refer to functional annotations; squares indicate proteins that were specifically modulated by ZIKV-NS4B that are significantly up- or downregulated during differentiation. $n = 4$ independent experiments, Bayesian statistical modelling. Unadjusted two-sided $P \leq 0.01$; $|\log_2(\text{fold change})| \geq 1$. **c**, Protein levels of cellular factors that were differentially modulated upon hNPC differentiation and ZIKV-NS4B-transduction. Ratios of protein abundance in naive hNPCs (differentiated/undifferentiated), differentiated hNPCs (ZIKV-NS4B/HCV-NS4B) or undifferentiated hNPCs (ZIKV-NS4B/HCV-NS4B). Coloured squares indicate functional annotations as in **b**. Up- or down- regulated proteins are displayed in shades of blue or green, respectively. Black circles indicate significant changes; sizes correspond to indicated P values. **d**, Total number of up- or downregulated class I phosphorylation sites significantly regulated by ZIKV infection 24, 48 and 72 h post-infection. $n = 4$, unadjusted two-sided $P \leq 0.01$; $|\log_2(\text{fold change})| \geq 1$. **e**, Selected cellular pathways and phosphorylation sites that are modified by ZIKV. Colours of protein nodes indicate measured ($\log_2$(fold change) of strongest phosphorylation site change) or inferred regulation of central proteins and substrates of canonical pathways. Individual phosphorylation sites are shown next to the corresponding molecules (colours indicate the highest $\log_2$(fold change) over time). Circles, newly identified phosphorylation sites; squares, known phosphorylation sites; framed squares, regulatory sites. Arrows and blunt-ended lines indicate positive or negative regulation of downstream proteins, respectively.

could be observed (Supplementary Tables 5, 8). This analysis highlights previously unreported ZIKV targets with known functions in neurogenesis. Among these targets, we note dephosphorylation of proteins, such as p38 MAPK, MARCKS (one of the main PKC substrates) and DPYSL2, activation of which positively modulates neurite outgrowth and brain development[21–23] (Fig. 2e and Extended Data Fig. 8e).

To assess the functional relevance of the newly identified host factors, we selected cellular targets regulated either at the proteome ($n = 26$), phosphoproteome ($n = 8$) or interactome ($n = 21$) level by ZIKV, and evaluated the consequences of virus infection on gene silencing (Fig. 3a, Extended Data Fig. 9a–c and Supplementary Table 9). This approach identified 17 host factors for which silencing significantly reduced ZIKV replication compared to non-targeting controls (Fig. 3a). In addition to confirming the essential role for recently described cellular factors in ZIKV replication, such as STT3A, MGGT1 and MSI1[24–26], we identified 14 additional host proteins that are required for ZIKV replication. Among these were proteins, the cellular abundance of which was strongly decreased upon NS4B expression and/or ZIKV infection in differentiating hNPCs (DOK3, XIRP2, YIPF4 and LMOD3), proteins that changed phosphorylation status upon ZIKV infection (LMO7, FXR1 and LARP7) and ZIKV-interacting proteins (LARP7 and LYAR, interacting with capsid, and BSG and CLN6, interacting

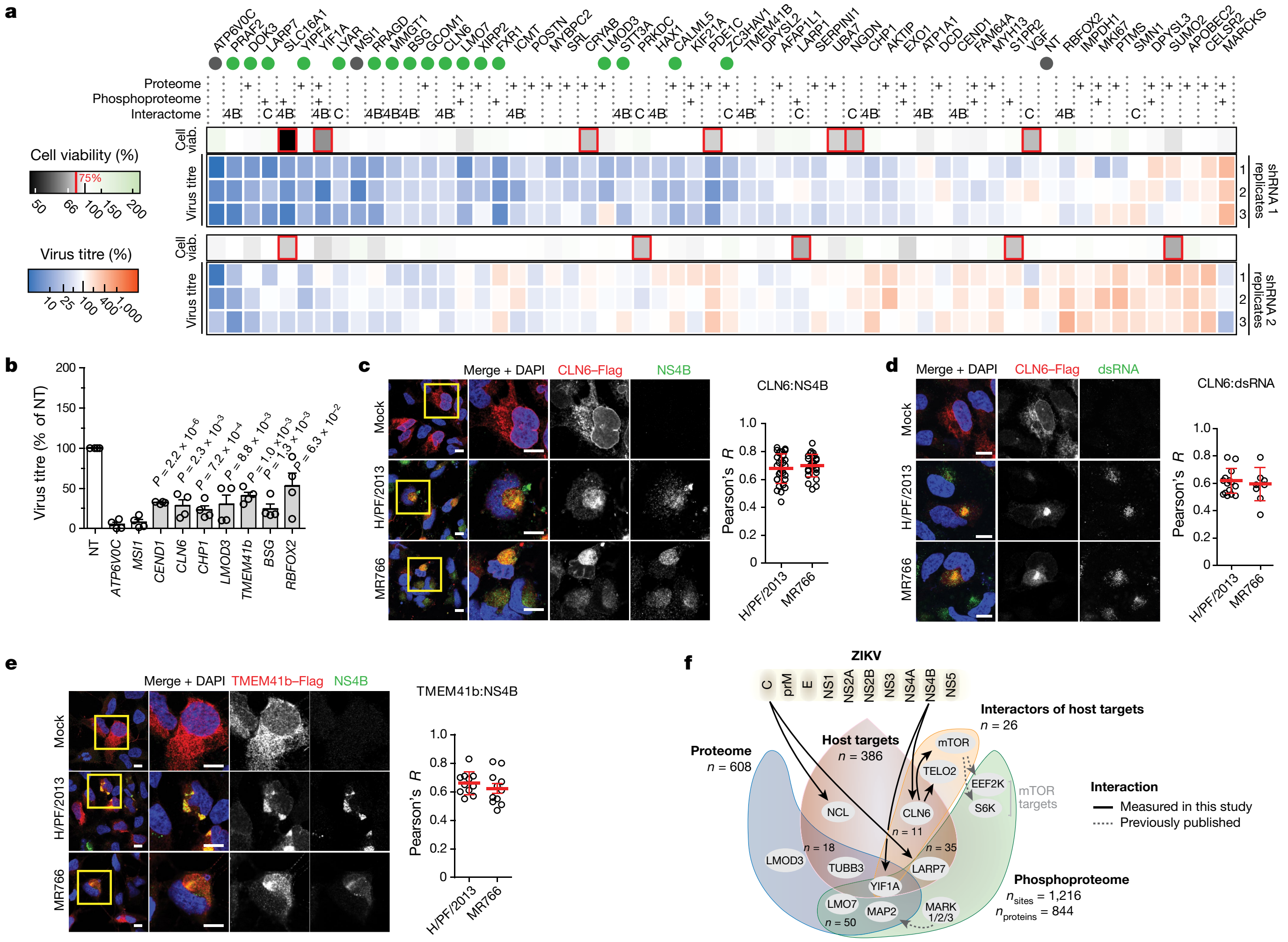


Fig. 3 | shRNA screen identifies novel ZIKV host-dependency factors. **a**, SK-N-BE2 cells were transduced with individual shRNAs that targeted genes selected from each of the proteomic surveys as indicated, and ZIKV replication was measured by plaque-forming unit (PFU) assay. Results from each biological repetition and shRNA are presented as a heat map. Non-targeting (NT) shRNA-normalized values of individual repetitions are shown in each row. Grey circles indicate non-targeting controls and shRNAs targeting known ZIKV host-dependency factors. Red squares indicate cytotoxic shRNAs (normalized median cell viability <75% of non-targeting controls; top rows). Green circles indicate significant hits (≥50% decrease in viral replication with both shRNAs each or ≥75% with any shRNA). **b**, Validation of silencing efficiency and gene knockdown effects on ZIKV replication on selected host factors from **a**. $n = 4$ independent experiments, normalized mean ± s.d.; two-sided unadjusted Student's *t*-test. **c**–**e**, Confocal microscopy of hNPCs transduced with CLN6- or TMEM41b-expressing lentiviruses (MOI = 3) and mock-infected (mock) or infected with ZIKV H/PF/2013 or MR766 (MOI = 5) for 24 h. Fixed cells were immunostained for dsRNA, ZIKV-NS4B and Flag-tagged proteins. Scale bars, 10 μm. $n = 2$ independent experiments. Representative images are shown. Graphs on the right display Pearson's correlation coefficients (mean ± s.d.; each dot represents a single cell). **f**, Intersection of ZIKV interactome, proteome and phosphoproteome datasets highlighting characteristic proteins perturbed at multiple levels. ZIKV infection disrupts phosphorylation (green) of its host targets (for example, LARP7) and along signalling pathways (for example, EEF2K and S6K via CLN6–mTOR–TELO2; or MAP2 via MARK1/2/3). Protein levels (blue) are also affected, carrying signatures of ZIKV-induced changes to vesicle transport (YIF1A) and neuronal differentiation pathways (MAP2). Solid arrows represent protein interactions measured in this study; grey dotted arrows represent published interactions. Numbers correspond to proteins significantly changed in a given dataset or intersection of datasets.

with NS4B) (Fig. 3a). Validation experiments confirmed that knockdown of CEND1, CLN6, CHP1, LMOD3, TMEM41b or BSG resulted in inhibition of ZIKV replication (Fig. 3b and Extended Data Fig. 9d, e). Moreover, infection of hNPCs with two different ZIKV isolates (H/PF/2013 and MR766) recruited CLN6 and TMEM41b to sites that were enriched in double-stranded (ds)RNA and NS4B (Fig. 3c–e).

We combined quantitative information at the proteome, phosphoproteome and interactome levels with publicly available interaction databases in order to obtain a holistic perspective on the effect of ZIKV on neuronal cells (Extended Data Fig. 10 and Supplementary Table 10). This orthogonal intersection can be used to infer hotspots that are targeted by ZIKV at multiple levels. For example, LARP7, which interacts specifically with capsid, is additionally hypophosphorylated, whereas LMO7 is simultaneously downregulated at the protein level and hyperphosphorylated upon ZIKV infection (Fig. 3f).

Our data provide valuable insights into how ZIKV exerts its action on candidate proteins that have previously been identified in genetic or drug-based screens. Among the recently identified ZIKV inhibitors[27] is digoxin, an inhibitor of the ATP1A1 Na^+/K^+ ATPase, which is identified here to interact specifically with NS4B. Fingolimod inhibits the sphingosine-1-phosphate receptor, which is persistently hypophosphorylated upon ZIKV infection. Mycophenolic acid impairs inosine-5-monophosphate dehydrogenase 1; the cellular abundance of this protein is markedly downregulated in ZIKV-infected cells. Although some of the ZIKV-binding partners that have been identified here may associate with orthologues of closely related flaviviruses (for example, dengue virus and West Nile virus), their role in ZIKV-related pathogenesis might in addition rely on the unique ability of ZIKV to cross the placental barrier and reach the developing brain[28].

Our study suggests that ZIKV evolved multiple mechanisms to usurp, exploit or perturb fundamental cellular processes, ultimately contributing to the broad spectrum of pathological abnormalities observed in humans. In addition, this study provides a resource for streamlining ZIKV research efforts and considering or excluding intracellular pathways for specific therapeutic interventions.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0484-5.

Received: 21 April 2017; Accepted: 20 July 2018;
Published online 3 September 2018.

1. Miner, J. J. & Diamond, M. S. Zika virus pathogenesis and tissue tropism. *Cell Host Microbe* **21**, 134–142 (2017).
2. Neufeldt, C. J., Cortese, M., Acosta, E. G. & Bartenschlager, R. Rewiring cellular networks by members of the *Flaviviridae* family. *Nat. Rev. Microbiol.* **16**, 125–142 (2018).
3. Chatel-Chaix, L. et al. Dengue virus perturbs mitochondrial morphodynamics to dampen innate immune responses. *Cell Host Microbe* **20**, 342–356 (2016).
4. Slomnicki, L. P. et al. Nucleolar enrichment of brain proteins with critical roles in human neurodevelopment. *Mol. Cell Proteomics* **15**, 2055–2075 (2016).
5. Li, H. et al. Ly-1 antibody reactive clone is an important nucleolar protein for control of self-renewal and differentiation in embryonic stem cells. *Stem Cells* **27**, 1244–1254 (2009).
6. Jung, M. Y., Lorenz, L. & Richter, J. D. Translational control by neuroguidin, a eukaryotic initiation factor 4E and CPEB binding protein. *Mol. Cell. Biol.* **26**, 4277–4287 (2006).
7. Liang, Q. et al. Zika virus NS4A and NS4B proteins deregulate Akt–mTOR signaling in human fetal neural stem cells to inhibit neurogenesis and induce autophagy. *Cell Stem Cell* **19**, 663–671 (2016).
8. Radke, J., Stenzel, W. & Goebel, H. H. Human NCL neuropathology. *Biochim. Biophys. Acta* **1852**, 2262–2266 (2015).
9. Ochrietor, J. D. & Linser, P. J. 5A11/Basigin gene products are necessary for proper maturation and function of the retina. *Dev. Neurosci.* **26**, 380–387 (2004).
10. Politis, P. K. et al. BM88/CEND1 coordinates cell cycle exit and differentiation of neuronal precursors. *Proc. Natl Acad. Sci. USA* **104**, 17861–17866 (2007).
11. Gehman, L. T. et al. The splicing regulator Rbfox2 is required for both cerebellar development and mature motor function. *Genes Dev.* **26**, 445–460 (2012).
12. Tarlungeanu, D. C. et al. Impaired amino acid transport at the blood brain barrier is a cause of autism spectrum disorder. *Cell* **167**, 1481–1494 (2016).
13. Souza, B. S. et al. Zika virus infection induces mitosis abnormalities and apoptotic cell death of human neural progenitor cells. *Sci. Rep.* **6**, 39775 (2016).
14. Tang, H. et al. Zika virus infects human cortical neural progenitors and attenuates their growth. *Cell Stem Cell* **18**, 587–590 (2016).
15. Watanabe, K. et al. A ROCK inhibitor permits survival of dissociated human embryonic stem cells. *Nat. Biotechnol.* **25**, 681–686 (2007).
16. Wojcechowskyj, J. A. et al. Quantitative phosphoproteomics reveals extensive cellular reprogramming during HIV-1 entry. *Cell Host Microbe* **13**, 613–623 (2013).
17. Söderholm, S. et al. Phosphoproteomics to characterize host response during influenza A virus infection of human macrophages. *Mol. Cell Proteomics* **15**, 3203–3219 (2016).
18. Humphrey, S. J., Azimifar, S. B. & Mann, M. High-throughput phosphoproteomics reveals in vivo insulin signaling dynamics. *Nat. Biotechnol.* **33**, 990–995 (2015).
19. Cao, B., Parnell, L. A., Diamond, M. S. & Mysorekar, I. U. Inhibition of autophagy limits vertical transmission of Zika virus in pregnant mice. *J. Exp. Med.* **214**, 2303–2313 (2017).
20. Ghouzzi, V. E. et al. ZIKA virus elicits P53 activation and genotoxic stress in human neural progenitors similar to mutations involved in severe forms of genetic microcephaly. *Cell Death Dis.* **7**, e2440 (2016).
21. Morooka, T. & Nishida, E. Requirement of p38 mitogen-activated protein kinase for neuronal differentiation in PC12 cells. *J. Biol. Chem.* **273**, 24285–24288 (1998).
22. Li, H., Chen, G., Zhou, B. & Duan, S. Actin filament assembly by myristoylated alanine-rich C kinase substrate-phosphatidylinositol-4,5-diphosphate signaling is critical for dendrite branching. *Mol. Biol. Cell* **19**, 4804–4813 (2008).
23. Yamashita, N. et al. Phosphorylation of CRMP2 (collapsin response mediator protein 2) is involved in proper dendritic field organization. *J. Neurosci.* **32**, 1360–1365 (2012).
24. Marceau, C. D. et al. Genetic dissection of *Flaviviridae* host factors through genome-scale CRISPR screens. *Nature* **535**, 159–163 (2016).
25. Savidis, G. et al. Identification of Zika virus and dengue virus dependency factors using functional genomics. *Cell Rep.* **16**, 232–246 (2016).
26. Chavali, P. L. et al. Neurodevelopmental protein Musashi-1 interacts with the Zika genome and promotes viral replication. *Science* **357**, 83–88 (2017).
27. Barrows, N. J. et al. A screen of FDA-approved drugs for inhibitors of Zika virus infection. *Cell Host Microbe* **20**, 259–270 (2016).
28. Shao, Q. et al. The African Zika virus MR-766 is more virulent and causes more severe brain damage than current Asian lineage and dengue virus. *Development* **144**, 4114–4124 (2017).

Acknowledgements We thank D. Mauceri, F. Sacco and L. Chatel-Chaix for discussions and R. Hornberger, I. Paron, K. Mayr and G. Sowa for technical assistance. The work in the authors' laboratory was funded by an ERC starting grant (StG 311339, iVIP), the Max-Planck free-floater program, the German Research Foundation (PI1084/2, PI1084/3, PI1084/4, TRR 237 and TRR179) and the Federal Ministry for Education and Research (ERA-Net grant ERASe) all to A.Pi. R.B. was supported by the Deutsche Forschungsgemeinschaft (BA1505/8-1). A. Pł. was supported by the Horizon 2020: Marie Skłodowska-Curie ETN "ANTIVIRALS" (grant agreement 642434 to R.B.). M.G. was supported by the DFG (SFB871) and the advanced ERC grant ChroNeuroRepair.

Reviewer information *Nature* thanks B. Berninger, I. M. Cristea, M. Evans and J. MacKenzie for their contribution to the peer review of this work.

Author contributions P.S. and A.Pi. conceived the study. P.S. performed most of the experiments. D.A.H. performed hNPC proteomics. K.D. performed hNPC cell culture. M.G. and R.B. intellectually contributed toward data interpretation. A.S. implemented the bioinformatic pipeline, statistical analysis and data integration. A. Pł. generated phosphoproteomic samples. M.C. generated and analysed immunofluorescence staining experiments. P.S. and A.Pi. wrote the manuscript with input from all authors.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0484-5.

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0484-5.

Reprints and permissions information is available at http://www.nature.com/reprints.

Correspondence and requests for materials should be addressed to P.S. or A.P.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Data reporting. No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Cell lines and reagents. HeLa S3 cells (CCL-2.2) and Vero E6 cells (CRL-1586) were purchased from ATCC. SK-N-BE2 cells were provided by R. Klein (MPI of Neurobiology, Munich). HEK293 cells were a gift from A. Bowie (Trinity College, Dublin). hNPCs were generated as described below. All cell lines were tested to be mycoplasma-free and their identity verified by STR profiling. Primary antibodies used in this study were: rabbit polyclonal anti-CEND1 (1:5,000 western blot; ab113076, Abcam), rabbit polyclonal anti-RBFOX2 (1:500 western blot; HPA006240, Sigma-Aldrich), rabbit polyclonal anti-CLN6 (1:500 western blot; SAB4502281, Sigma-Aldrich), rabbit polyclonal anti-CHP1 (1:500 western blot; PA5-29876, Thermo-Fisher), rabbit polyclonal anti-BSG (1:2,000 western blot; HPA036048, Sigma-Aldrich), rabbit polyclonal anti-TMEM41b (1:250 western blot; HPA014946, Sigma-Aldrich), mouse monoclonal anti-dsRNA J2 (1:400 immunofluorescence; Scicons), mouse IgG2bκ anti-haemagglutinin (HA) (1:500 immunoflorescence; 12CA5, Sigma-Aldrich), rabbit polyclonal anti-DENV-NS4B (dengue virus protein NS4B) (1:100 IFA; 1:500 western blot; GTX124250; Genetex; this antibody cross-reacts with ZIKV-NS4B, as validated by immunodetection of ZIKV-NS4B in cells transduced with HA-tagged NS4B or ZIKV-infected cell lysates (Extended Data Fig. 4b)). For detection of NS4B and capsid upon ZIKV infection by western blotting, the rabbit polyclonal anti-ZIKV-NS4B (1:1,000 western blot; GTX133311; Genetex) and anti-ZIKV capsid (1:1,000 western blot; GTX133317, Genetex) were used. For detection of phosphorylated and total protein abundance by western blotting, mouse monoclonal anti-gammaH2Ax (ab22551, Abcam) and the following rabbit monoclonal antibodies were used (Cell Signaling Technology): anti-MARCKS (D88D11) and anti-phosphorylated MARCKS (D13E4), rabbit monoclonal anti-mTOR (7C10) and anti-phosphorylated mTOR (D9C2), p38–MAPK (D13E1) and anti-phosphorylated p38–MAPK (D3F9), S6 ribosomal protein (5G10) and anti-phosphorylated S6 (D57.2.2E). Where indicated, to increase the basal phosphorylation of certain cellular proteins, 12-*O*-tetradecanoylphorbol-13-acetate (TPA, Cell Signaling Technologies) or DMSO, were added to the culture medium (32 nM), 15 min before collection of the cell lysates for western blot analysis. For characterization of hNPCs by immunofluorescence, the following antibodies were used: mouse anti-GFAP (G3893, Sigma-Aldrich), rat anti-Ki-67 (14-5698, EBioscience), mouse anti-MAP2 (M4403, clone HM-2, Sigma-Aldrich), mouse anti-nestin (MAB5326, clone 10C2, Millipore) and rabbit anti-Sox2 (ab137385, Abcam).

Generation, maintenance and differentiation of hNPCs. Human induced pluripotent stem cells (iPSCs) generated from healthy patient fibroblasts[29] were maintained in mTeSRTM 1 medium (STEMCELL Technologies) on Geltrex (ThermoFischer Scientific) plates in the absence of feeders. The iPSCs were regularly confirmed to be mycoplasma-negative. To generate hNPCs, iPSC colonies were collected with a cell scraper to ensure that relatively big colonies remained and these were cultured in suspension to form embryoid bodies in DMEM/F-12, 0.5% N2 supplement (Gibco), 5 μM dorsomorphin (Sigma-Aldrich), 10 μM SB431542 (Sigma-Aldrich) and 10 μM Rho-associated kinase inhibitor (ROCK; Y-27632, StemCell Technologies). Embryoid bodies were grown for five days and the medium was changed every other day before gentle trituration and attachment on poly-ornithine–laminin-coated dishes in NPC medium, which consisted of DMEM/F-12, 0.5% N2, 1% B27 supplement with 10 ng ml^{-1} bFGF (Peprotech 100-18B-50) and 10 μM ROCK inhibitor. Three to four days after plating the embryoid bodies, neural rosettes were manually isolated, dissociated mechanically to single cells and plated on fresh poly-ornithine–laminin-coated dishes in NPC medium with 10 μM ROCK inhibitor. On the next day, the medium was changed to NPC medium without ROCK and NPCs were propagated for up to 12 passages by splitting with accutase when 70–80% confluency was reached. For short-term differentiation experiments (Fig. 3c–e and Extended Data Fig. 4), 2 × 10^5 cells were seeded in six-well plates, transduced with empty lentiviruses (NT), lentiviruses expressing HCV- or ZIKV-NS4B or infected with an MOI of 0.01 of ZIKV H/PF/2013. After 48 h, differentiation was induced for five days by growth factor withdrawal and addition of 10 μM ROCK inhibitor, while undifferentiated cells were kept in NPC medium in the presence of bFGF. In both cases, the medium was replaced every other day. Cell pellets were collected, snap-frozen and processed for LC–MS/MS analysis as described below.

Virus strains, virus stocks preparation and generation of lentiviruses. ZIKV strains H/PF/2013 and MR766 were obtained from the European Virus Archive (EVAg, France). Virus stocks were passaged once on VeroE6 cells and virus-containing cell culture supernatants or conditioned medium from uninfected cells were collected from day 3 to 8 post-infection. Supernatants were filtered through a 0.45-μM pore-size filter and stored at −70 °C. Titres of infectious virus were determined by plaque assay as described below. For infection of hNPCs, virus stocks were diluted in hNPC medium and cells were inoculated for 2 h at 37 °C. Inocula were removed and fresh hNPC medium (basal or differentiation) was added. Mock cells were inoculated with equal amounts of conditioned medium.

For the generation of lentiviruses that expressed the individual HA-tagged ZIKV viral proteins or Flag-tagged host proteins, synthetic cDNAs (sequences based on recombinant ZIKV strain FSS13025, Asian lineage) were obtained in pUDC57 vectors and the fusion gene was inserted into the lentiviral pWPI expression plasmid. The lentivirus that expresses the HA-tagged HCV-NS4B of has previously been described[30]. Lentivirus stocks were produced and titrated following standard procedures. Nucleotide sequences of each construct are available from the corresponding authors upon request.

Lentiviral shRNA library production and shRNA screen. From the MISSION TRC lentiviral library (Sigma-Aldrich), 54 MS hits and 4 controls were selected and shRNA were produced as follows. Two different shRNA-expressing lentiviruses per gene (116 shRNAs in total; Supplementary Table 10) were individually produced in HEK 293T cells (8 × 10^5) that were plated one day before transfection. Transfections were performed as previously described[4] using the packaging plasmids psPAX2 and pMD-VSV-G (provided by D. Trono) and shRNA-encoding pLKO-puro plasmids. Viruses were collected at 48 and 72 h post-transfection, clarified by centrifugation and pooled before freezing. Controls included the non-targeting shRNA-encoding plasmids SHC001 and SHC002 (NT1 and NT2) as well as the known virus restriction factors ATP6V0C and Musashi1. For quality-control purposes, approximately 25% of random samples were used to measure lentiviral titres. For screening purposes, 10,000 SK-N-BE2 cells per well were seeded in 96-well plates in 90 μl of complete DMEM and 24 h later cells were transduced with 30 μl of each lentivirus (MOI ≈ 3) in triplicate wells/replica and three independent biological replicates. Three days later, cells were infected with an MOI of 0.1 of ZIKV H/PF/2013. Subsequently, 48 h post-infection, cell supernatants were used for titration by PFU assay. Cell viability was measured by a resazurin-based cell-viability assay, on plates which were transduced with shRNAs in parallel but were not infected with ZIKV. In brief, 50 μg ml^{-1} resazurin was added to each well of a 96-well plate and incubated for 30 min at 37 °C, followed by measurement of fluorescence (535/590 nm) using an Infinite 200 PRO series microplate reader (Tecan). The normalization for the virus titres and fluorescence measurements were performed in two steps. First, the measurements in each replicate experiment were adjusted by the median factor of the replicate (the median of $X_{g,i}$/median($X_{g,1}$, $X_{g,2}$, $X_{g,3}$) across all the shRNAs ($g = 1,2,...,116$), where $X_{g,i}$ is the titre of *g*th shRNA in *i*th replicate). Second, the adjusted measurements were normalized, such that the average signal of non-targeting controls was 100%. The following criteria were applied to select hits: ≥50% decrease in median viral titres with two shRNAs or ≥75% if only one shRNA was efficient; cell viability ≥75% of non-targeting controls. Validation experiments for selected shRNAs (Fig. 3b and Extended Data Fig. 9d, e) were performed in 24-well plates using the same overall protocol.

Virus titration by plaque assay. Confluent monolayers of VeroE6 cells were infected with serial tenfold dilutions of virus supernatants for 2 h at 37 °C. The inoculum was removed and replaced with serum-free MEM (Gibco, Life Technologies) containing 1.5% carboxymethylcellulose (Sigma-Aldrich). Four days post-infection, cells were fixed for 2 h at room temperature with formaldehyde directly added to the medium to a final concentration of 5%. Fixed cells were washed extensively with water before staining with H_2O containing 1% crystal violet and 10% ethanol for 30 min. After rinsing with water, the number of plaques was counted and virus titres were calculated.

Sample preparation for deep-proteomic and phosphoproteomic analyses. For proteomic and phosphoproteomic analysis, SK-N-BE2 cells (10^7 cells per biological replicate per condition; 4 biological replicates per condition), were mock-infected or infected with ZIKV H/PF/2013 at an MOI of 3. Cell pellets were collected after 24, 48 and 72 h, lysed in 1 ml of lysis buffer (10 mM Tris-HCl (pH 7.5), 4% SDS and 0.05 mM DTT supplemented with complete protease and phosphatase inhibitor cocktails (Roche)), boiled 10 min at 98 °C, and sonicated (4 °C for 15 min, or until a homogeneous suspension was formed). Clarified protein lysates were precipitated with acetone, and normalized protein mixtures resuspended in 500 μl TFE digestion buffer. Protein digestion was performed by adding 1:100 (protein:enzyme) trypsin and LysC with rapid agitation (2,000 r.p.m.) overnight at 37 °C. An aliquot of the peptide mixtures was used for the determination of the total proteome (10%), while the remaining peptide mixture (90%) was processed for phosphopeptide enrichment using the EasyPhos protocol. For proteomic analysis of primary cells, hNPCs were seeded in six-well plates (2 × 10^5 cells per well) and either mock-infected, infected with an MOI of 0.01 of ZIKV or transduced with lentiviruses expressing HA-tagged ZIKV-NS4B or HCV-NS4B (MOI = 3). Two days after infection or lentivirus transduction, cells were left in hNPC proliferation medium or neuronal differentiation was initiated by culturing the cells in hNPC differentiation medium. Five days later, cell pellets were collected and processed for label-free LC–MS/MS analysis as described previously[31]. Each condition was performed in quadruplicate biological replicates.

Deep-proteome and phosphoproteome LC–MS/MS analysis and data processing. Peptide mixtures were separated on a 50-cm reversed-phase column (diameter of 75 μm packed in-house with ReproSil-Pur C18-AQ 1.9 μm resin (Dr. Maisch GmbH)) as previously described[32].

Raw MS files were processed within the MaxQuant environment (version 1.5.7) using the integrated Andromeda search engine with false discovery rate (FDR) ≤ 0.01 at the protein, peptide and modification level. Proteome and phosphoproteome files were assigned to separate search parameter groups. The search included fixed modifications for carbamidomethyl (C) and variable modifications for oxidized methionine (M), acetylation (protein N-term), and phospho (STY) for phosphoproteome files. Peptides with at least six amino acids were considered for identification, and 'match between runs' was enabled with a matching time window of 0.7 min to transfer MS1 identifications between runs. Peptides and proteins were identified using a UniProt FASTA database from human (UniprotKB release 2015_08 including isoforms and unreviewed sequences) and ZIKV virus polyprotein corresponding to the H/PF/2013 strain (NCBI GenBank KJ776791.2; individual viral cleavage products were manually annotated).

Sample preparation for affinity purification of ZIKV proteins and cellular proteins. For the determination of the ZIKV interactome, four independent affinity purifications were performed for each ZIKV HA-tagged viral protein. SK-N-BE2 cells were transduced with an MOI of 3 of each lentivirus (8×10^6 cells per dish) and 72 h later cells were scraped into 1 ml of lysis buffer (50 mM Tris (pH 8), 150 mM NaCl, 0.5% NP-40, cOmplete protease inhibitor cocktail, Roche) and HA-affinity purifications were performed as described previously[33]. In brief, clarified cell lysates were incubated with anti-HA-specific beads for 3 h at 4 °C, and non-specifically bound proteins removed by three washes with lysis buffer and five washes with washing buffer (50 mM Tris (pH 8), 150 mM NaCl). Bound proteins were denatured by incubation in 20 μl guanidinium chloride buffer (600 mM GdmCl, 1 mM TCEP, 4 mM CAA, 100 mM Tris-HCl (pH 8)). After digestion with 1 μg LysC (WAKO Chemicals USA) at room temperature for 3 h, the suspension was diluted in 100 mM Tris-HCl (pH 8), and the protein solution was digested with trypsin (Promega) overnight at room temperature. Peptides were purified on stage tips with three C18 Empore filter discs (3M) and analysed by mass spectrometry as previously described[34]. For the analysis of CLN6 and other host-factor interactomes, SK-N-BE2 cells were transduced with the corresponding Flag-tagged lentiviruses and 72 h later cell lysates were processed as described above, using agarose anti-Flag M2 beads (Sigma-Aldrich).

Data processing of AP–LC–MS/MS samples. Raw mass-spectrometry data were processed with MaxQuant (software version 1.5.3) using the built-in Andromeda search engine to search against the human proteome (UniprotKB release 2015_08 including isoforms and unreviewed sequences) plus ZIKV virus polyprotein (ZIKV strain FSS13025, Asian lineage; NCBI GenBank KU955593.1 with individual viral cleavage products manually annotated), Gaussia luciferase (Q9BLZ2), HCV-NS4B (E7ELX2) sequences, and the label-free quantification algorithm as described previously[35]. In MaxQuant, carbamidomethylation was set as fixed and methionine oxidation and N-terminal acetylation as variable modifications, using an initial mass tolerance of 6 p.p.m. for the precursor ion and 0.5 Da for the fragment ions. Search results were filtered with a FDR of 0.01 for peptide and protein identifications.

Statistical analysis of MS data. MaxQuant output files (proteinGroups.txt and 'Phospho (STY) Sites.txt' for proteome and phosphoproteome data, respectively) were processed by a combination of in-house R (version 3.3), Julia (version 0.5) and Stan (version 2.14) scripts.

To account for the variation in the analysed biological sample amounts and MS performance, we inferred the normalizing multipliers for each MS run so that the normalized LFQ values of the representative set of proteins (around 500 randomly values were selected from the ones quantified in most of the normalized samples) do not change between the samples. First, this procedure was applied to normalize replicate MS runs within each condition, and then the normalization multipliers were further adjusted to normalize the conditions between each other. For the analysis of proteome changes the LFQ intensities were fit to a Bayesian generalized linear model (GLM) for each protein group (raw intensities of each phosphosite for phosphoproteome data) individually. The model used horseshoe priors for the treatment effects, and the natural logarithm as a linking function for protein intensities. Additional parameters with horseshoe priors were introduced to model the biological variation within replicate samples. The MS measurements were set to follow the Laplace distribution. Technical MS runs from both LTQ-Orbitrap XL and Orbitrap Q Exactive HF were used to estimate the signal-to-noise ratio and sensitivity of protein quantification across the whole dynamic range of protein intensities. We used the obtained technical variation models to define the scale parameter of the Laplace distribution for each individual MS measurement as well as scale and location parameters of the logistic distribution for missing MS measurements. To analyse the hNPC proteome, we used the following GLM (in R GLM formula notation): $\log(\text{intensity}) \propto 1 + \text{differentiation} \times \text{treatment}$ with horseshoe prior parameters $\tau_1 = 1$ and $\tau_2 = 0.25$ for model effects and variation in biological replicates, respectively. For the SK-N-BE2-cell dataset the corresponding statistical model was $\log(\text{intensity}) \propto 1 + (\text{after}_{24\,\text{h}} + \text{after}_{48\,\text{h}}) \times \text{treatment}$ with $\tau_1 = 2$ and $\tau_2 = 0.1$ horseshoe prior parameters for proteomic data and $\tau_1 = 1$ and $\tau_2 = 0.25$ for phosphorylation sites (phosphosites). The GLM model for AP–LC–MS/MS experiments was $\log(\text{intensity}) \propto 1 + \text{bait}$, with $\tau_1 = 1$ and $\tau_2 = 0.3$ for ZIKV baits, and $\tau_1 = 2$ and $\tau_2 = 0.1$ when host proteins were used as baits.

The posterior distributions of model parameters were inferred using the Hamiltonian Markov chain Monte Carlo (MCMC) method (Stan version 2.14), using eight parallel MCMC chains, 2,000 warmup and 2,000 sampling iterations taking every fourth sample from each chain. The statistical significance of a hypothesis that given model variable x is positive given data Y (one-sided P value), that is, the probability that a random sample from the posterior distribution of x would be non-positive ($P(x \leq 0|Y)$), was calculated by approximating the posterior distribution from the MCMC samples using Gaussian kernel with Silverman's rule-of-thumb bandwidth. For the $x \neq 0$ hypothesis, the two-sided P value was defined as $\min(P(x \leq 0|Y), P(x \geq 0|Y))$. The following criteria were used for reporting significantly regulated proteins and phosphosites: $P \leq 0.01$ and $|\text{median}(\log_2(\text{fold change}))| \geq 1$ (Supplementary Tables 2, 5, 9).

To filter for specific interactions, the inferred posterior distributions of protein intensities in each viral bait of the ZIKV AP–LC–MS/MS dataset were compared with the average distribution of the same protein in all other baits excluding capsid protein (owing to its highly hydrophobic character and multiple subcellular localizations). The interaction was considered specific, if the one-sided P value ($P(x_{\text{bait}} \leq x_{\text{background}}|Y)$, where x_{bait} is the protein intensity in the pull-downs of a given bait, and $x_{\text{background}}$ is its intensity in the pull-down experiments of the other baits) was below 5×10^{-5} for capsid bait (since there were no other baits with the same biochemical properties) and 5×10^{-2} for the other baits, and median protein enrichment against the background was more than 5.56-fold ($= 2^{2.5}$) for high-confidence and twofold for the lower-confidence set of ZIKV protein interactions. The same filtering strategy was applied to AP–LC–MS/MS experiments of selected host targets; the P-value cutoff was 0.02 and $\log_2(\text{fold change}) \geq 1.5$. For the representation of networks, subcellular localization of cellular proteins was extracted from Gene Ontology (version 2016-Sep-21).

Merging of datasets. The integration of all AP–LC–MS/MS, proteomic and phosphoproteomic data was done by matching protein groups of individual datasets to each other using in-house scripts. The protein groups were matched if they shared a common protein accession code ('Majority Protein ACs' column in MaxQuant output). In ambiguous situations (for example, one protein group of dataset A shares different protein sequences with several protein groups of dataset B), the priority was given to the reviewed protein sequences and canonical isoforms.

Integrating published protein–protein interactions, kinase–substrate relations and regulatory sites data. For overlaying our data with published protein–protein interactions, we used the IntAct interactions database (version 2017.04.08), IntAct complexes collection (version 2017.04.08) and CORUM protein complexes database (version 2017.03.15). The interaction (physical association) was included if both interacting partners were significantly enriched in the same ZIKV bait. For published AP–MS interactions, we additionally required that one of the interacting partners was a bait. To integrate AP–LC–MS/MS datasets with proteomic and phosphoproteomic screens, only published direct interactions (including kinase–substrate relationships) were considered.

For phosphoproteomic analysis kinase–substrate relations and regulatory sites were extracted from PhosphositePlus[36] (kinase_substrate_dataset from 21 March 2017 and regulatory_sites from 18 April 2017, respectively).

Ingenuity pathway analysis. Proteome and phosphoproteome changes along with unchanged proteins or sites, respectively, were submitted to the core ingenuity pathway analysis (IPA). The following cut-offs were set for differentially expressed proteins: $P \leq 0.01$ and $|\log_2(\text{fold change})| \geq 1.5$. Ingenuity knowledge base was used as a reference dataset, while only experimentally observed findings were used for confidence filtering. No filters were applied at the level of species, tissues or cell lines. Input datasets were used for the functional analysis to identify biological functions and canonical pathways that were most significant. Right-tailed Fisher's exact tests were used to calculate P values. Biological functions or pathways were considered significant when unadjusted $P \leq 0.05$. Corresponding Benjamini–Hochberg-adjusted P values were calculated and are presented in the Supplementary Tables.

Co-immunoprecipitation assays. For validation of ZIKV-NS4B interacting proteins, SK-N-BE2 cells were seeded into 15-cm^2 dishes (8×10^6 cells per dish), transduced with lentiviruses expressing ZIKV, DENV or HCV HA-tagged viral proteins or conditioned cell-culture medium. After 72 h of transduction, cell monolayers were scraped into 1 ml lysis buffer (50 mM Tris-HCl (pH 8.0), 0.5% NP-40, 150 mM NaCl and protease inhibitor cocktail (cOmplete, Roche)), and clarified lysates were processed for immunoprecipitation as previously described[33]. Eluted proteins were further analysed by western blot as specified in the figure legends. For reciprocal co-immunoprecipitation experiments, SK-N-BE2 cells were transduced with lentiviruses expressing each Flag-tagged host protein and three days later were infected with ZIKV at an MOI of 3. After 72 h of infection, cell monolayers were processed for immunoprecipitation with anti-Flag M2 agarose beads (Sigma-Aldrich) as previously described[33].

Immunofluorescence analysis. hNPCs (10^5 cells per well) seeded on poly-ornithine–laminin-coated glass slides were transduced with lentiviruses (MOI = 3) expressing Flag-tagged proteins or empty control for 1 h at 37 °C. The next day, the medium was changed and 48 h post-transduction, cells were infected with ZIKV MR766 or H/PF/2013 at an MOI of 5 for 1 h at 37 °C. After 24 h, cells were fixed with 4% PFA, permeabilized with 0.1% (vol/vol) Triton X-100 in PBS and unspecific binding sites were blocked with PBS containing 5% BSA for 1 h at room temperature, and subjected to immunofluorescence staining using anti-Flag antibody, anti-DENV-NS4B or J2 anti-dsRNA. Secondary staining was performed with donkey anti-mouse Alexa568-conjugated and goat anti-rabbit Alexa 488-conjugated antibodies. Nuclear DNA was stained with 4′,6-diamidino-2-phenylindole (DAPI) (Molecular Probes). Coverslips were mounted in fluoromount-G mounting medium (Southern Biotechnology Associates). Fluorescence images were acquired with a Leica SP8 inverted confocal microscope using a 63× Plan-Apo NA 1.4 objective. Image analysis was performed using FIJI (http://fiji.sc/wiki/index.php/Fiji) and Pearson's correlation coefficients were calculated using the Coloc2 plugin[37].

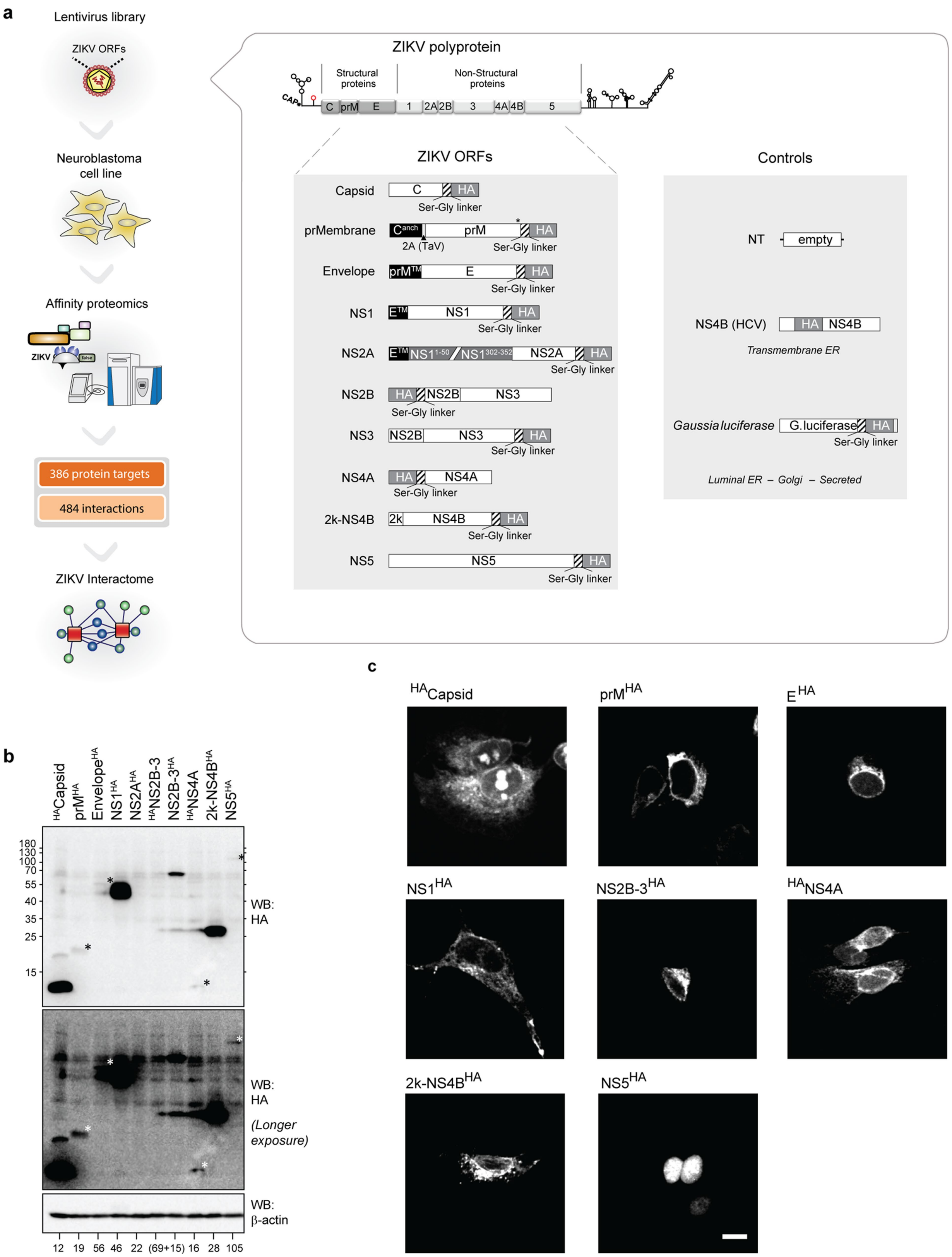
Accession numbers. UniprotKB accession codes for all protein groups identified by MS are provided in each respective Supplementary Table. The corresponding sequences were retrieved from UniprotKB (human; release 2015_08, including isoforms and unreviewed sequences). cDNA sequences corresponding to individual ZIKV open-reading frames used for generation of ZIKV lentivirus library (KU955593.1) or related to the wild-type ZIKV H/PF/2013 strain (KJ776791.2) were extracted from GenBank. Protein sequences of Gaussia luciferase (Q9BLZ2) and HCV-NS4B protein (E7ELX2) were extracted from UniprotKB.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Code availability. The analysis scripts written in R (version 3.3), Julia (version 0.5) and Stan (version 2.14) are available from the corresponding authors upon request.

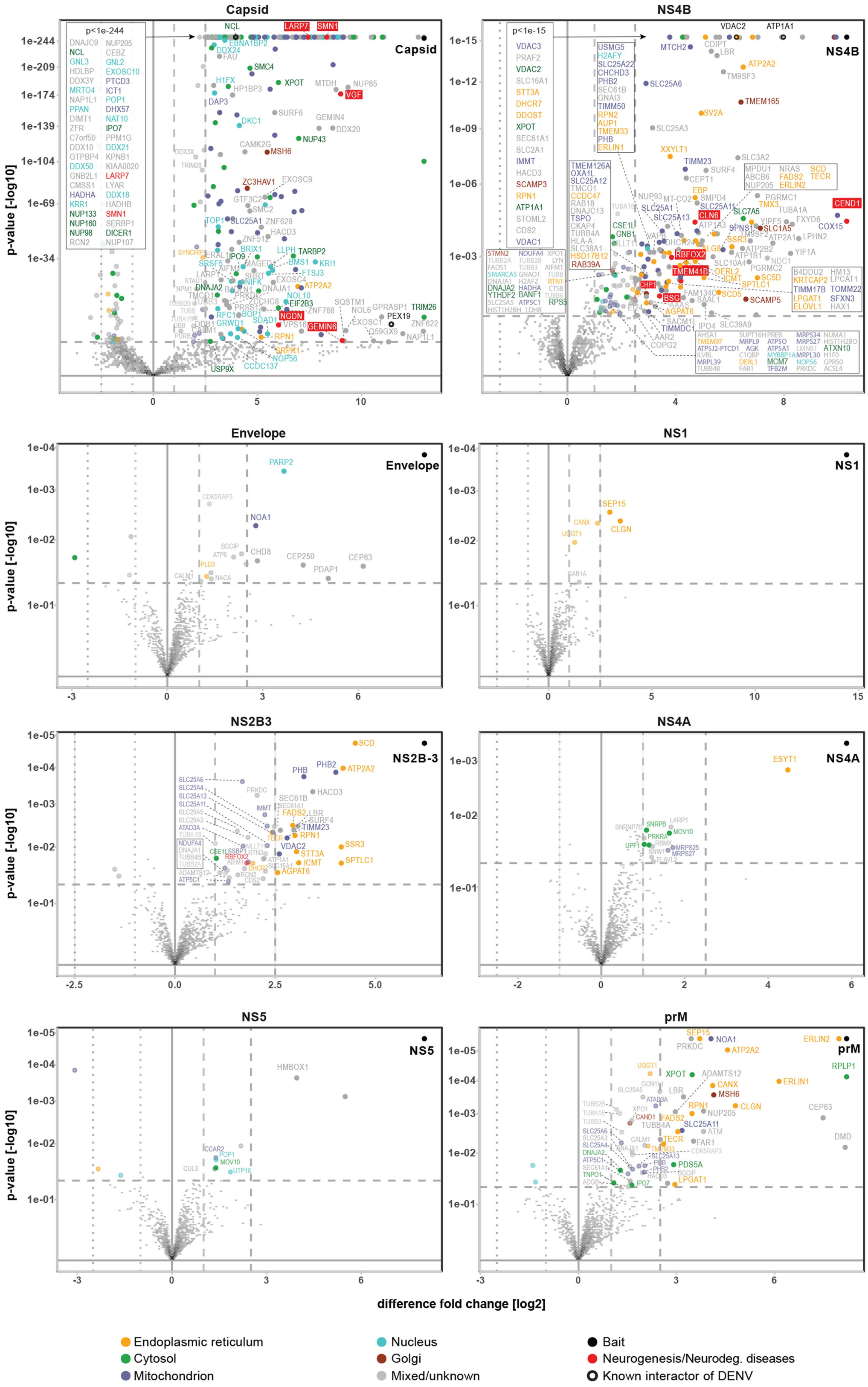
Data availability. The mass-spectrometry-based proteomics data were deposited at the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository with the following dataset identifiers: PXD009551, PXD009557, PXD009560 and PXD009561. The protein interactions from this publication have been submitted to the IMEx (http://www.imexconsortium.org) consortium through IntAct[38] with the identifier IM-26452.

29. Havlicek, S. et al. Gene dosage-dependent rescue of HSP neurite defects in SPG4 patients' neurons. *Human Molecular Genetics* **23**, 2527–2541 (2014).
30. Paul, D., Bartenschlager, R. & McCormick, C. The predominant species of nonstructural protein 4B in hepatitis C virus-replicating cells is not palmitoylated. *J. Gen. Virol.* **96**, 1696–1701 (2015).
31. Rappsilber, J., Mann, M. & Ishihama, Y. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat. Protoc.* **2**, 1896–1906 (2007).
32. Steger, M. et al. Phosphoproteomics reveals that Parkinson's disease kinase LRRK2 regulates a subset of Rab GTPases. *eLife* **5**, e12813 (2016).
33. Scaturro, P., Cortese, M., Chatel-Chaix, L., Fischl, W. & Bartenschlager, R. Dengue virus non-structural protein 1 modulates infectious particle production via interaction with the structural proteins. *PLoS Pathog.* **11**, e1005277 (2015).
34. Gebhardt, A. et al. mRNA export through an additional cap-binding complex consisting of NCBP1 and NCBP3. *Nat. Commun.* **6**, 8192 (2015).
35. Hubner, N. C. et al. Quantitative proteomics combined with BAC TransgeneOmics reveals in vivo protein interactions. *J. Cell Biol.* **189**, 739–754 (2010).
36. Hornbeck, P. V. et al. PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res.* **43**, D512–D520 (2015).
37. Schindelin, J. et al. Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).
38. Orchard, S. et al. The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.* **42**, D358–D363 (2014).

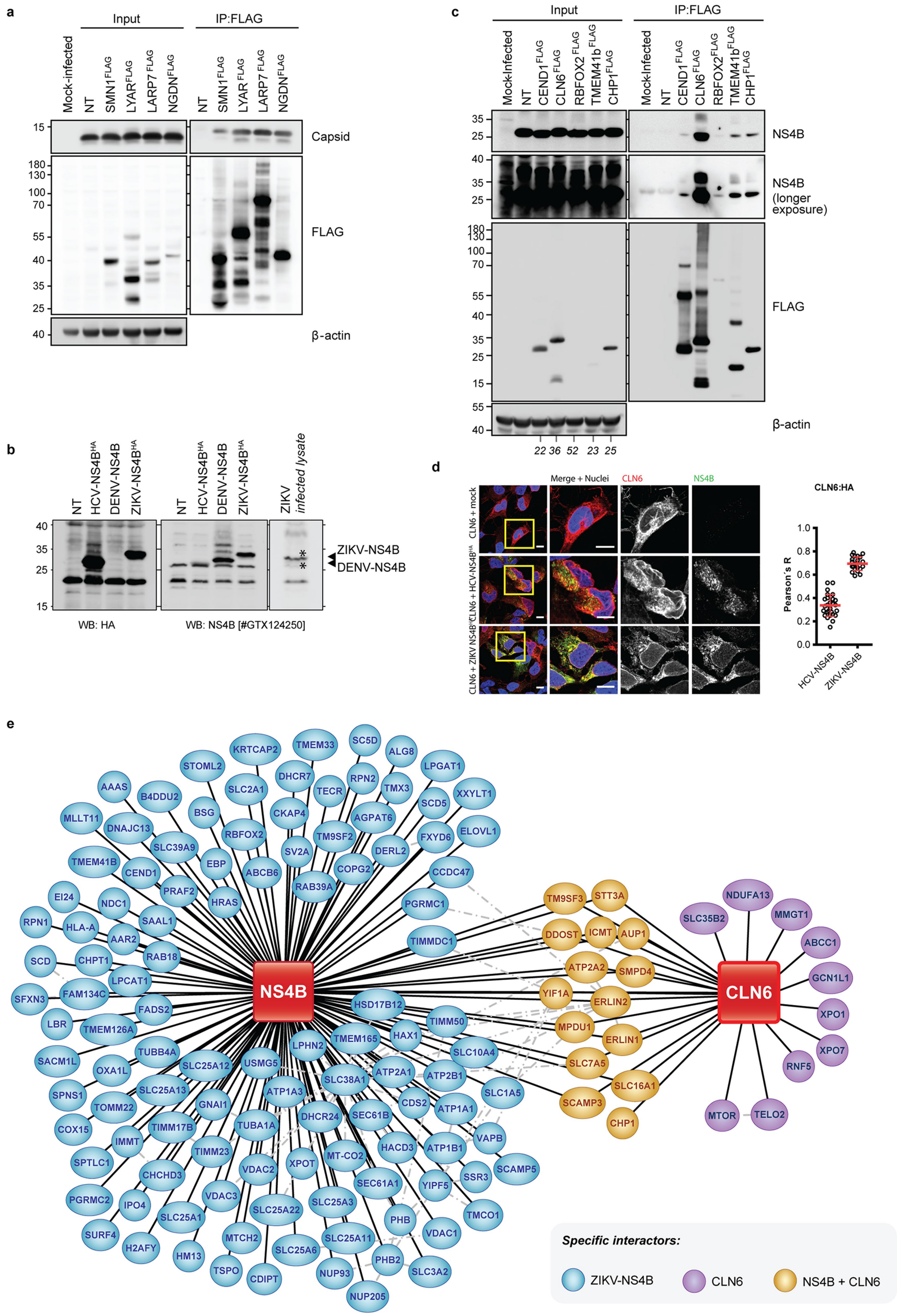

a
Lentivirus library
ZIKV ORFs
Neuroblastoma cell line
Affinity proteomics
ZIKV
386 protein targets
484 interactions
ZIKV Interactome
ZIKV polyprotein
Structural proteins
Non-Structural proteins
CAP
C prM E 1 2A 2B 3 4A 4B 5
ZIKV ORFs
Capsid
C HA
Ser-Gly linker
prMembrane
C^anch prM HA
2A (TaV)
Ser-Gly linker
Envelope
prM^TM E HA
Ser-Gly linker
NS1
E^TM NS1 HA
Ser-Gly linker
NS2A
E^TM NS1^1-50 NS1^302-352 NS2A HA
Ser-Gly linker
NS2B
HA NS2B NS3
Ser-Gly linker
NS3
NS2B NS3 HA
Ser-Gly linker
NS4A
HA NS4A
Ser-Gly linker
2k-NS4B
2k NS4B HA
Ser-Gly linker
NS5
NS5 HA
Ser-Gly linker
Controls
NT
empty
NS4B (HCV)
HA NS4B
Transmembrane ER
Gaussia luciferase
G.luciferase HA
Ser-Gly linker
Luminal ER – Golgi – Secreted
b
^HACapsid
prM^HA
Envelope^HA
NS1^HA
NS2A^HA
^HANS2B-3
NS2B-3^HA
^HANS4A
2k-NS4B^HA
NS5^HA
180
130
100
70
55
40
35
25
15
WB: HA
WB: HA
(Longer exposure)
WB: β-actin
12 19 56 46 22 (69+15) 16 28 105
c
^HACapsid
prM^HA
E^HA
NS1^HA
NS2B-3^HA
^HANS4A
2k-NS4B^HA
NS5^HA

Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Design rationale, expression and subcellular localization of HA-tagged ZIKV proteins. a, Schematic representation of ZIKV HA-tagged viral proteins and controls used in this study and experimental set-up used to generate the ZIKV cellular interactome. The full-length ZIKV genome is shown at the top, with the 5′ and 3′ untranslated regions depicted with their putative secondary structures. Polyprotein cleavage products are separated by vertical lines and labelled as specified. The individual open reading frames (ORFs) of each of the ZIKV proteins were fused with an HA epitope either at their N terminus or C terminus as depicted. Numbers above each sequence refer to the nucleotide sequence of the ZIKV isolate FSS13025 used as reference (GeneBank: KU955593.1). To ensure the correct subcellular localization and protein topology, the sequence of the capsid anchor (C^{anch}), the transmembrane domain of prM (prM^{TM}) and the transmembrane domain of Envelope (E^{TM}) were fused at the N termini of prM, envelope and NS1, respectively. Additionally, given the complex topology of NS2A, the transmembrane domain of envelope, and a fusion protein encoding the first and the last 50 amino acids of NS1 was fused at its N terminus. As additional controls, an empty pWPI-lentiviral vector (NT) or HA-tagged non-structural protein 4B of hepatitis C virus (NS4B-HCV) and Gaussia luciferase (G. luciferase) were included to monitor for non-specific binding to the anti-HA beads as well as organelle-dependent enrichment artefacts in AP–LC–MS/MS analysis. **b**, Intracellular levels of ZIKV proteins in lentivirus-transduced SK-N-BE2 cells. SK-N-BE2 cells were transduced with lentiviruses encoding each of the ZIKV proteins at an MOI of 3, and 72 h later cell lysates that had been clarified by centrifugation were used for western blotting against HA and β-actin. Numbers on the left refer to molecular weight standards given in kDa. Asterisks mark proteins expressed at a lower level, which can be observed upon longer exposure of the membrane. Note that C-terminally tagged NS2A and N-terminally tagged NS2B-3 could not be detected, whereas the C-terminally tagged NS2B-3 fusion protein was expressed at the expected molecular weight, and was therefore used for further experiments in this study. The bottom lane indicates the predicted molecular weight of each viral protein in kDa. $n = 2$ independent experiments. A representative blot is shown. **c**, Subcellular localization of ZIKV proteins in lentivirus-transduced SK-N-BE2 cells. SK-N-BE2 cells were seeded in 24-well plates and transduced as described above. After 72 h of transduction, viral proteins were detected using rabbit anti-HA antibody and Alexa-Fluor488-conjugated secondary antibody and imaged by confocal microscopy. Scale bar, 10 μm. $n = 1$ experiment.

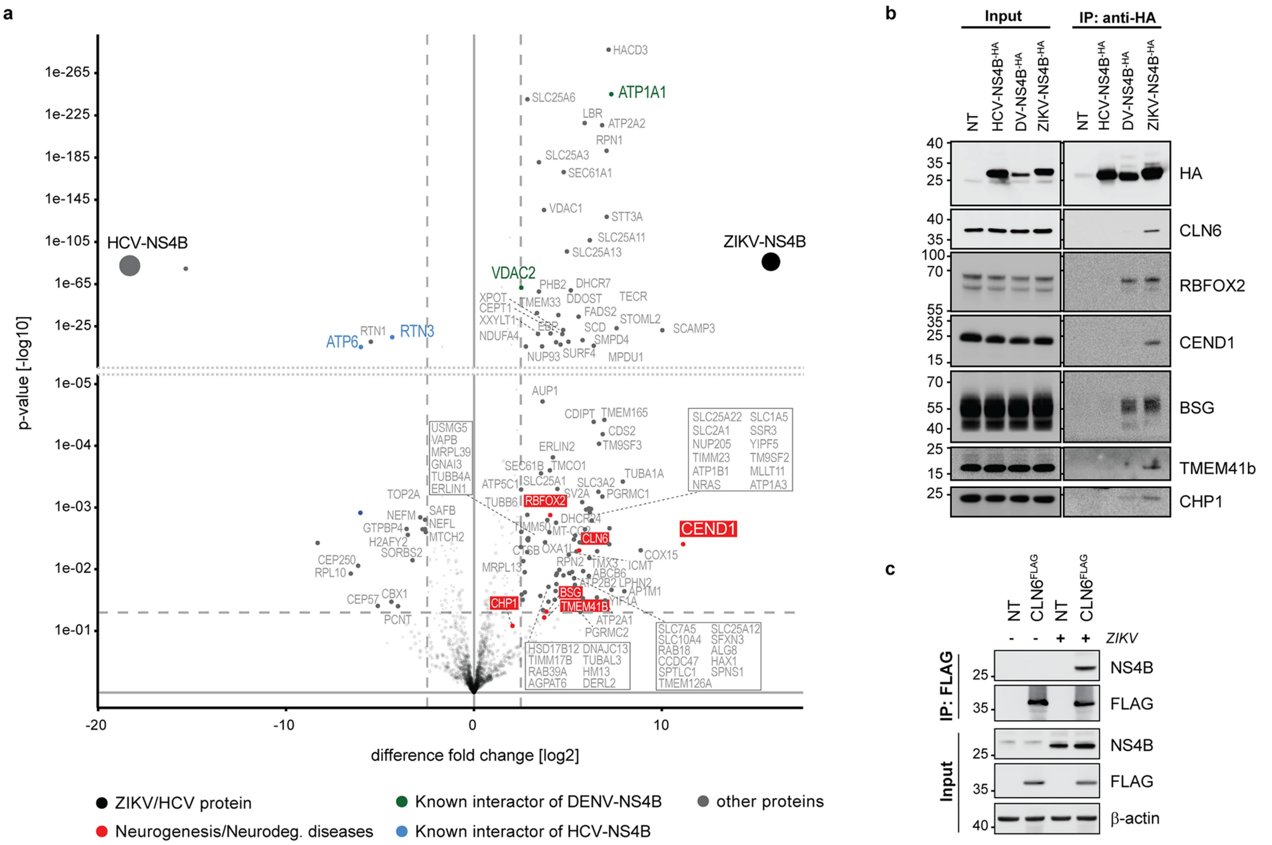


Extended Data Fig. 2 | Volcano plots of individual ZIKV-interacting proteins. Volcano plots of label-free AP–LC–MS/MS of individual ZIKV proteins. Each volcano plot displays all identified proteins (median $\log_2$(fold change) of bait-specific protein enrichment in comparison to the background plotted against the corresponding $-\log_{10}(P \text{ value})$. Dotted grey lines represent the $\log_2$(fold change) and P value cut-offs used. High-confidence interacting proteins ($\log_2(\text{fold change}) \geq 2.5$, unadjusted one-sided $P \leq 0.01$) are coloured according to their subcellular localization. In the case of capsid, due to its highly hydrophobic character and multiple subcellular localizations, a more stringent P value cut-off was used (unadjusted one-sided $P \leq 5 \times 10^{-5}$) and ribosomal proteins were not individually labelled. The respective viral bait of each AP is shown in black, black circles represent known DENV target proteins while proteins with roles in neurogenesis, neurodegenerative diseases or infertility are labelled in red. $n = 4$ independent experiments, significance testing (P values) results from Bayesian statistical modelling (see Methods, 'Statistical analysis of MS data').

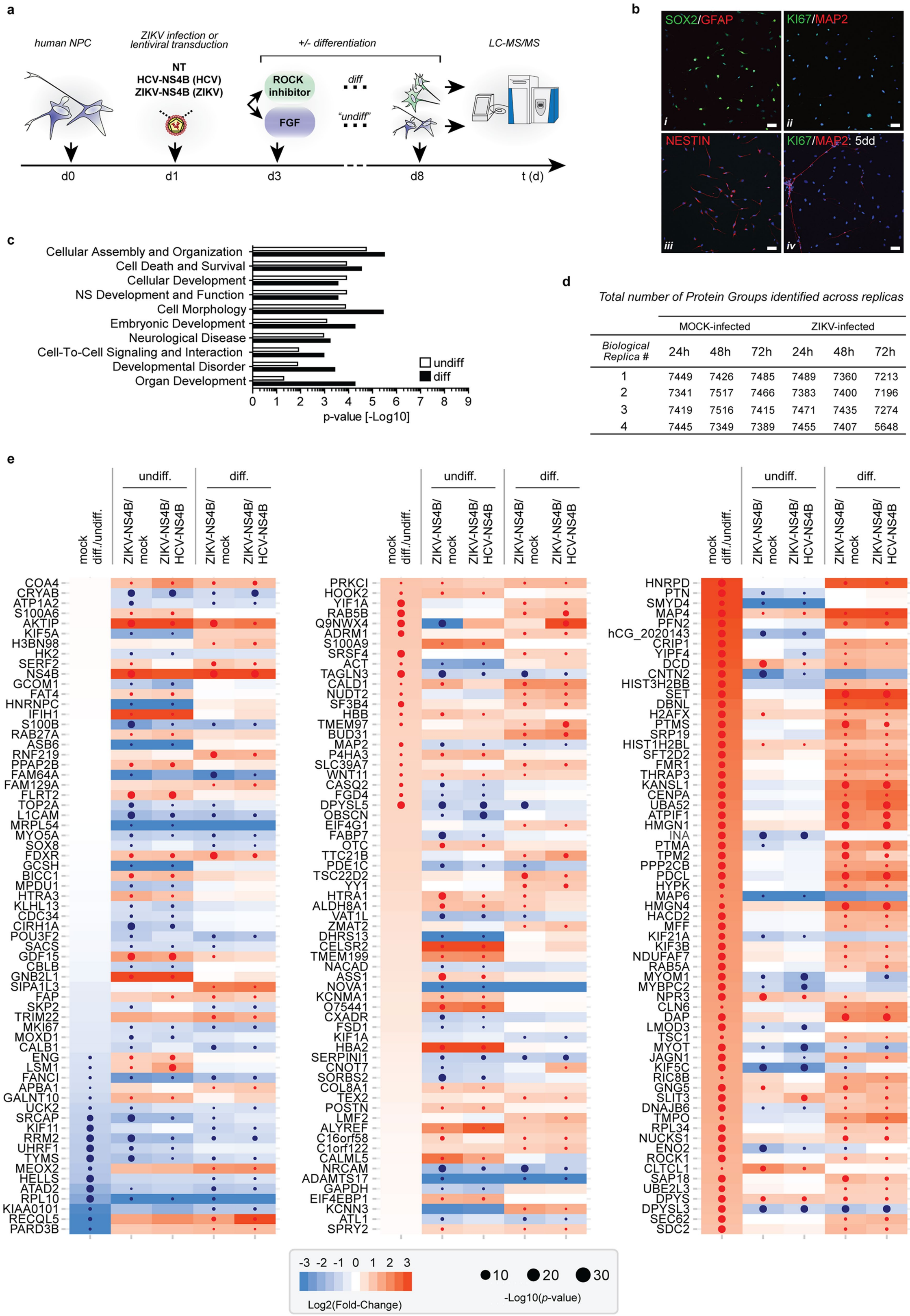


Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Validation of NS4B-interacting proteins upon ZIKV infection. a, Reciprocal co-immunoprecipitation of NS4B with Flag-tagged host factors. SK-N-BE2 cells transiently transduced with empty lentiviruses (NT) or lentiviruses expressing Flag-tagged CEND1, CLN6, TMEM41b, RBFOX2 and CHP1 were mock-infected or infected with ZIKV H/PF/2013 (MOI = 1) and three days later Flag-immunoprecipitated proteins were probed with a ZIKV-NS4B-specific antibody. The bottom lane indicates the predicted molecular weights of each viral protein in kDa. *n* = 3 independent experiments. Representative blots are shown. **b**, Validation of anti-DENV-NS4B-specific antibody for immunodetection of ZIKV-NS4B. SK-N-BE2 cells were transiently transduced with empty lentiviruses (NT) or lentiviruses expressing HCV-NS4B–HA, DENV-NS4B or ZIKV-NS4B–HA, and 72 h later were probed consecutively with mouse anti-HA (left) or rabbit anti-DENV-NS4B (right) specific antibodies. ZIKV-infected SK-N-BE2 cell lysates, shown on the right side, were included as control. Asterisks indicate unspecific bands. *n* = 1 experiment. **c**, Reciprocal co-immunoprecipitation of capsid with Flag-tagged host factors. SK-N-BE2 cells transiently transduced with empty lentiviruses (NT) or lentiviruses expressing Flag-tagged SMN1, LYAR, LARP7 and NGDN were infected as described above and Flag-immunoprecipitated proteins probed with a ZIKV capsid-specific antibody. *n* = 2 independent experiments. Representative blots are shown. **d**, hNPCs were transduced with lentiviruses expressing Flag-tagged CLN6 alone (mock) or co-transduced with lentiviruses expressing HCV-NS4B–HA or ZIKV-NS4B–HA (MOI = 3), stained with Flag- and HA-specific antibodies and imaged by confocal microscopy. Boxed areas are enlarged on the right. Data are mean ± s.d. of Pearson's correlation coefficients (each dot represent a single cell). Scale bars, 10 μm. *n* = 2 independent experiments. **e**, Network representation of CLN6-interacting cellular proteins and comparison with the ZIKV-NS4B interactors. AP–LC–MS/MS analysis of Flag-tagged CLN6 in SK-N-BE2 cells revealed novel CLN6-specific interacting proteins (displayed in magenta; including TELO2 and mTOR) as well as shared cellular interactors with ZIKV-NS4B (displayed in yellow, including ICMT and STT3A). Solid lines represent specific interactors identified by AP–MS/MS analysis ($\log_2$(fold change) ≥ 1.5, unadjusted one-sided $P \leq 0.02$), grey dotted lines represent previously published protein–protein interactions. *n* = 4 independent experiments, significance testing results from Bayesian statistical modelling.



Extended Data Fig. 4 | Volcano plots of ZIKV-NS4B and HCV-NS4B-interacting proteins. a, Volcano plot comparing the specificity of protein enrichment in ZIKV-NS4B versus HCV-NS4B AP (median $\log_2$(fold change) plotted against $-\log_{10}$(unadjusted *P*) for ZIKV-NS4B versus HCV-NS4B). Specific NS4B-interacting proteins are labelled as described in Extended Data Fig. 2. Previously reported interactors of HCV-NS4B, or the ZIKV closely related DENV-NS4B are shown in blue and green, respectively. $n = 4$ independent experiments, significant testing results from Bayesian statistical modelling. **b**, Co-immunoprecipitation of ZIKV-NS4B–HA with endogenous host proteins. Cell lysates of SK-N-BE2 cells transduced with empty lentiviruses (NT) or lentivirus-expressing HA-tagged HCV-, DENV- or ZIKV-NS4B proteins were used for HA-immunoaffinity purification and probed with the indicated antibodies. $n = 3$ independent experiments. Representative blots are shown. **c**, Reciprocal co-immunoprecipitation of NS4B with Flag-tagged CLN6. SK-N-BE2 cells transiently transduced with empty lentiviruses (NT) or lentivirus-expressing Flag-tagged CLN6 were infected with ZIKV (H/PF/2013) and Flag-immunoprecipitated proteins probed with NS4B-specific antibodies. $n = 3$ independent experiments. Representative blots are shown.

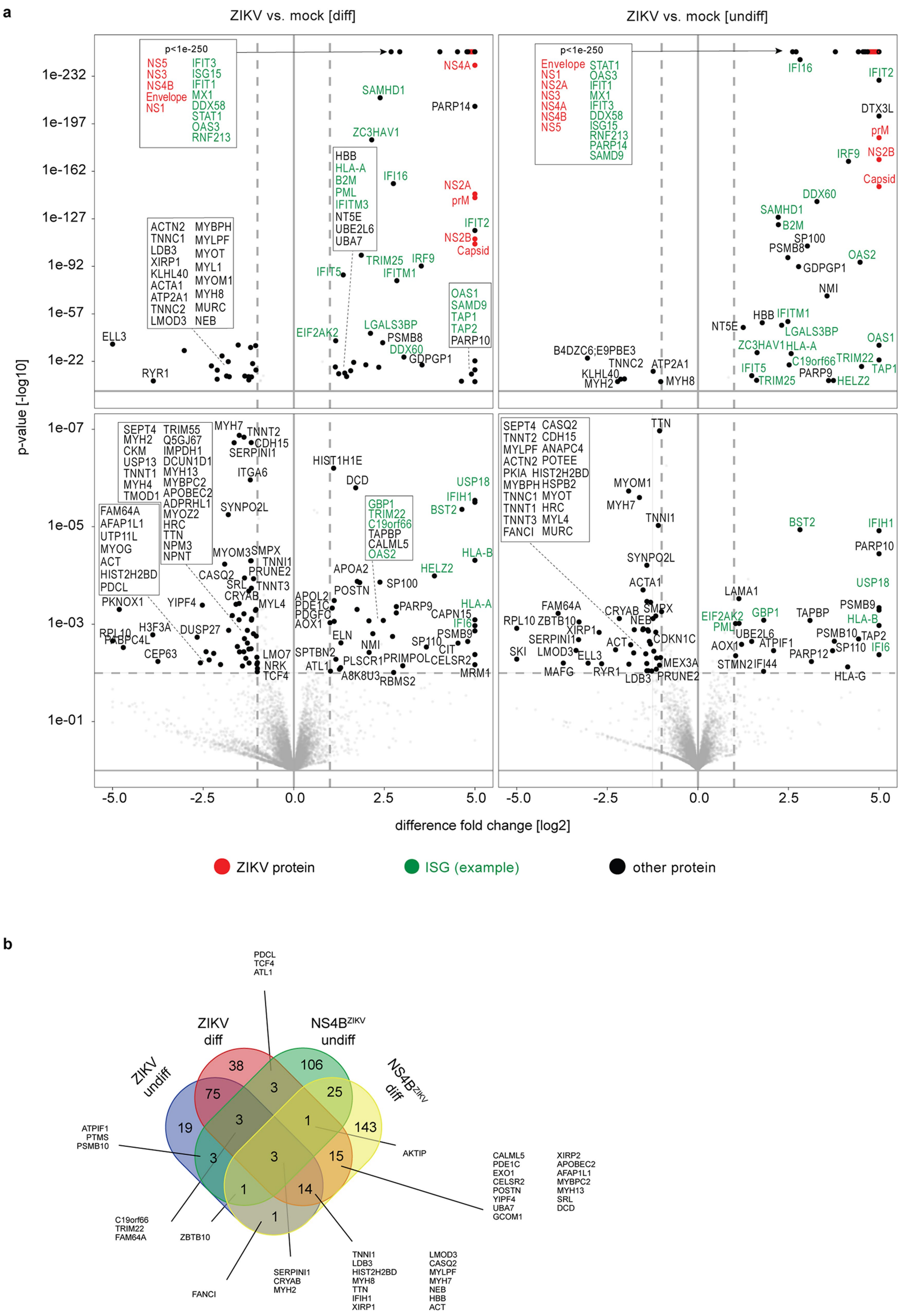


Total number of Protein Groups identified across replicas

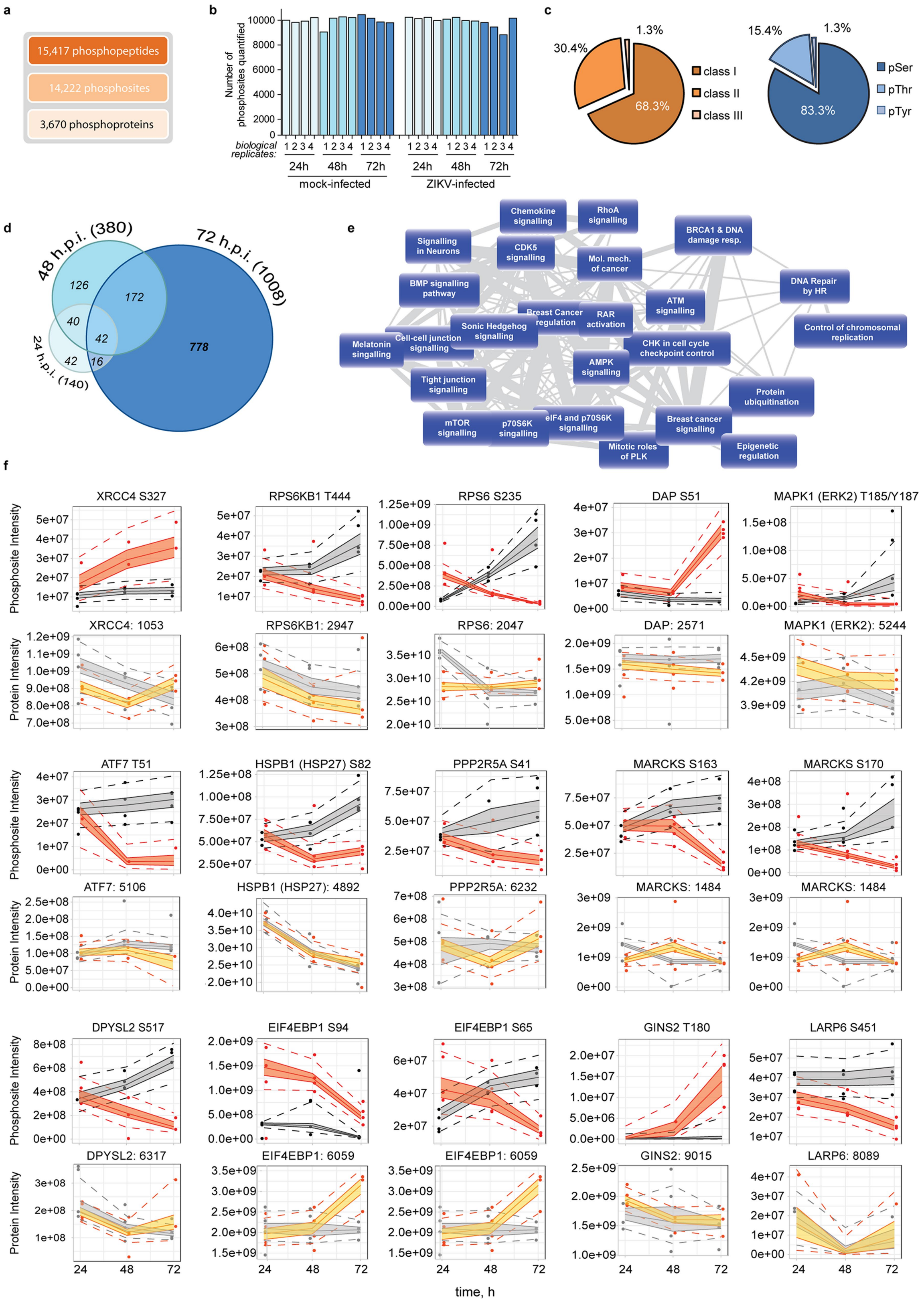
Biological Replica #	MOCK-infected 24h	MOCK-infected 48h	MOCK-infected 72h	ZIKV-infected 24h	ZIKV-infected 48h	ZIKV-infected 72h
1	7449	7426	7485	7489	7360	7213
2	7341	7517	7466	7383	7400	7196
3	7419	7516	7415	7471	7435	7274
4	7445	7349	7389	7455	7407	5648

Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | Effects of ZIKV infection or NS4B transduction on undifferentiated or differentiated hNPCs. a, Experimental approach used for hNPC differentiation. hNPCs were transduced with empty (NT), HCV-NS4B–HA- or ZIKV-NS4B–HA-expressing lentiviruses (MOI = 3), mock-infected or infected with ZIKV (H/PF/2013) (MOI = 0.01). After 48 h, differentiation was induced for five days by growth factor withdrawal and addition of 10 μM ROCK inhibitor, while undifferentiated cells were kept in NPC medium in the presence of bFGF. Cell pellets were then used for quantitative label-free LC–MS/MS-based profiling of the global proteome. **b**, Characterization of hNPCs used for proteomic analysis. Undifferentiated hNPCs kept in the presence of bFGF were stained with (i) SOX2- and GFAP-, (ii) Ki-67- and MAP2-, and (iii) nestin-specific antibodies. Alternatively, hNPCs were differentiated for five days in the presence of a ROCK inhibitor and were stained with (iv) Ki-67- and MAP2-specific antibodies. Under proliferating conditions, hNPCs expressed Ki-67, SOX2 and nestin, confirming the maintenance of their proliferative state in the presence of bFGF. No neuron-specific protein expression (MAP2) or astrocyte marker expression (GFAP) was detected under these conditions. After differentiation for five days in vitro in the presence of ROCK inhibitor, the neuronal marker MAP2 was upregulated whereas the proliferation marker Ki-67 was downregulated, suggesting commitment towards the neuronal lineage. Scale bars, 20 μm. $n = 3$ independent experiments. Representative images are shown. **c**, Cellular processes enriched in NS4B-ZIKV changes in proliferating or differentiating hNPCs. Right-tailed Fisher's exact test, Benjamini–Hochberg-corrected *P* values are shown (see Supplementary Table 3). **d**, Total number of proteins identified across biological replicates and conditions. **e**, Heat map of all significant changes occurring upon proliferation and differentiation conditions in NS4B-ZIKV-transduced cells in comparison to NS4B-HCV-tranduced cells and mock transduction. Unadjusted $P \leq 0.01$, $|\log_2(\text{fold change})| \geq 1.0$; significance testing results from Bayesian statistical modelling. **c**–**e**, $n = 4$ independent experiments.

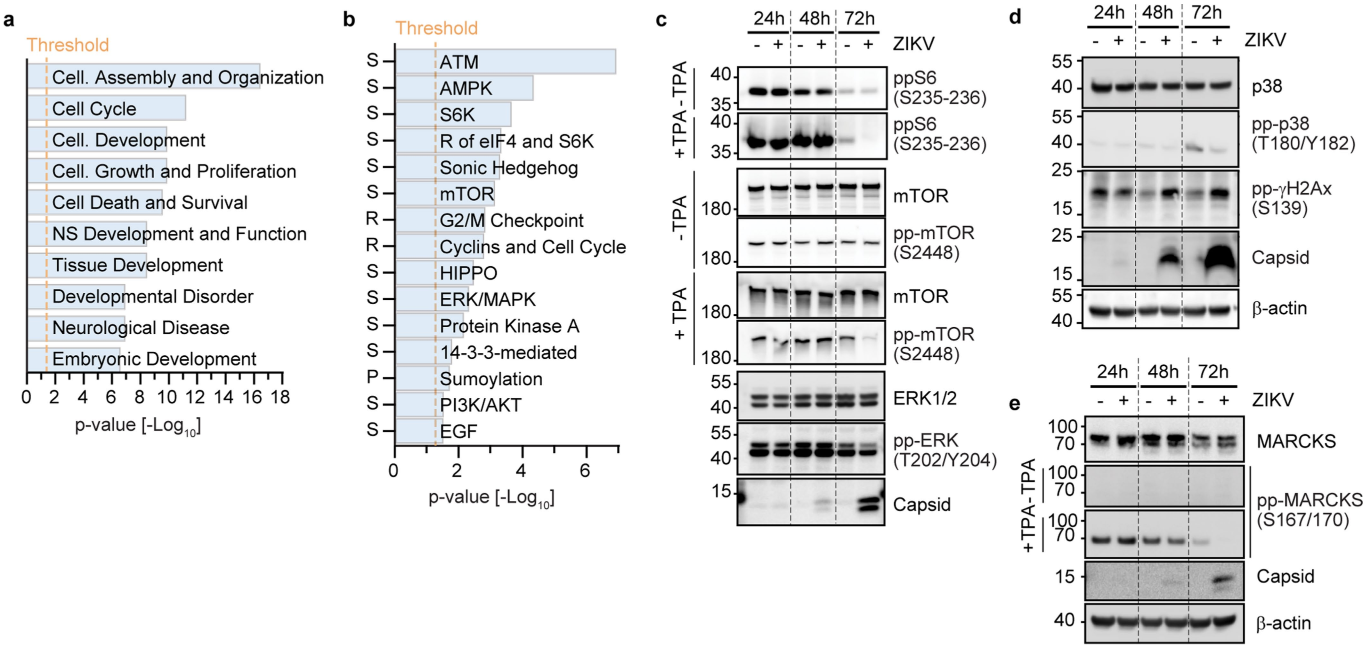


Extended Data Fig. 6 | Global proteomic analysis of ZIKV-specific effects in differentiating or proliferating hNPCs. a, hNPCs were mock-infected or infected with ZIKV (H/PF/2013) (MOI = 0.01) and cultured under proliferating or differentiating conditions as described in Extended Data Fig. 5a. Volcano plots display proteins that were significantly up- or downregulated by ZIKV-infection in differentiated (left) or undifferentiated (right) hNPCs. Viral proteins are labelled in red. Proteins with functions in antiviral immunity are labelled in green. Significance cut-offs are indicated by the dashed grey lines. *n* = 4 independent experiments. Bayesian statistical modelling, $|\log_2(\text{fold change})| \geq 1$; unadjusted two-sided $P \leq 0.01$. **b,** Venn diagram displaying the total number of significantly up- or downregulated proteins in every experimental condition ($|\log_2(\text{fold change})| \geq 1$; $P < 0.01$) and the total number and gene names of significantly modulated proteins similarly regulated by ZIKV infection or ZIKV-NS4B transduction upon differentiation (diff) or proliferation (undiff). *n* = 4 independent experiments.

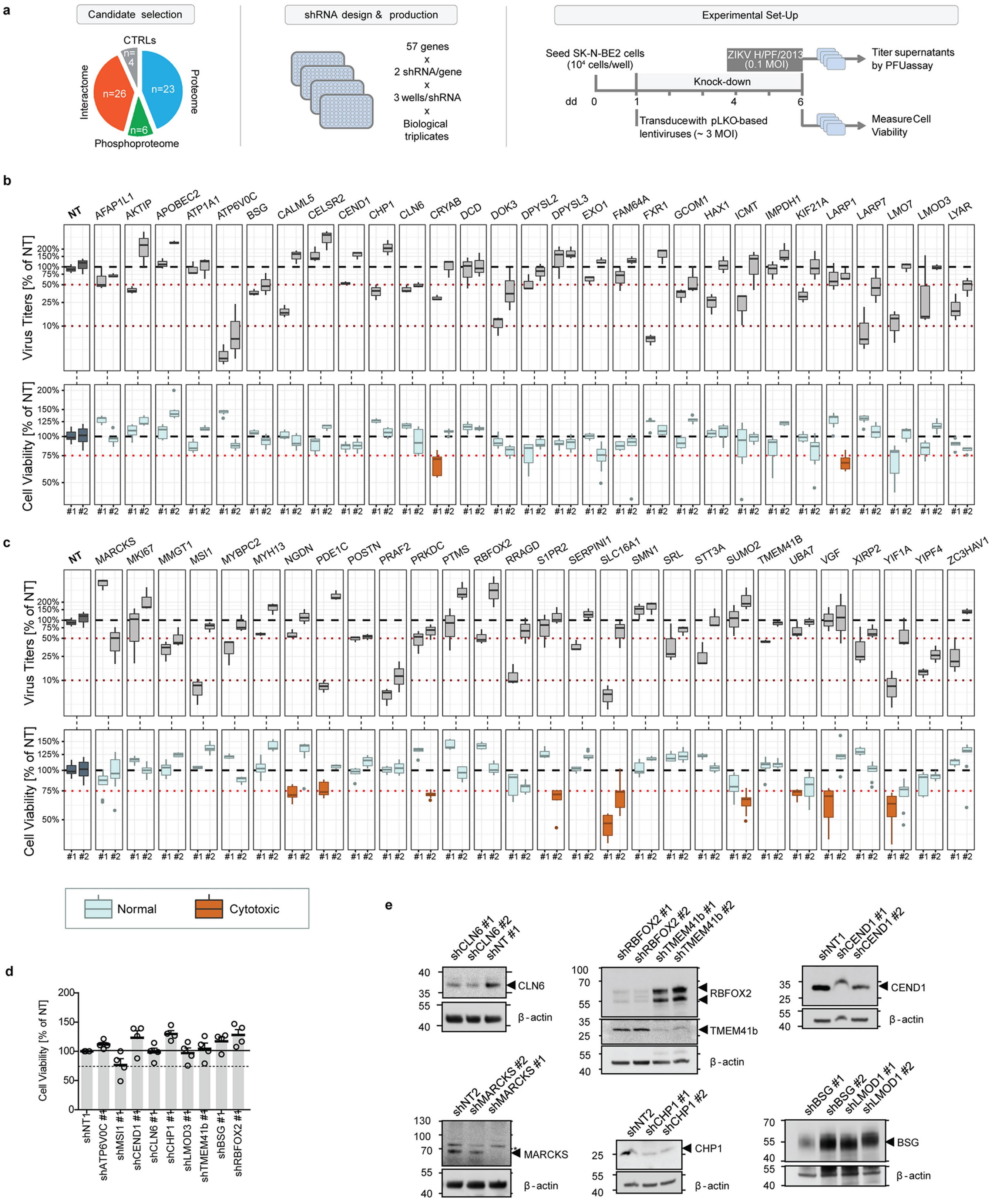


Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Time-resolved phosphoproteomic analysis of ZIKV-infected SK-N-BE2 cells. a, Total number of phosphopeptides, phosphosites and phosphoproteins identified in this study. **b**, Total number of phosphosites identified across biological replicates and conditions ($n = 4$). **c**, Relative percentage of class I, II and III phosphosites (localization probability $>$0.75) and phosphorylation of specific Ser, Thr and Tyr residues. **d**, Venn diagram of significantly changing phosphosites at 24, 48 and 72 h post-infection (h.p.i.). **e**, Network analysis of cellular processes significantly modulated by ZIKV infection at phosphorylation level (IPA). Nodes are canonical pathways identified with Fisher's exact test; edges are shared proteins between the pathways ($n = 4$ independent experiments, right-tailed Fisher's exact test, Benjamini–Hochberg-adjusted $P \leq 0.05$). **f**, Profile plots of significantly changing phosphosites, mapped to pathways in Fig. 3d, and their corresponding total protein levels at 24, 48 and 72 h after ZIKV infection (orange for phosphorylation levels, yellow for protein levels) or mock treatment (grey). Numbers next to protein names refer to MaxQuant protein group IDs (Supplementary Table 5). Points are normalized intensities of individual replicates, solid line is median, filled area corresponds to 25–75 percentiles, dashed lines mark 2.5–97.5 percentiles of the posterior distribution. $n = 4$ independent experiments; Bayesian statistical modelling.

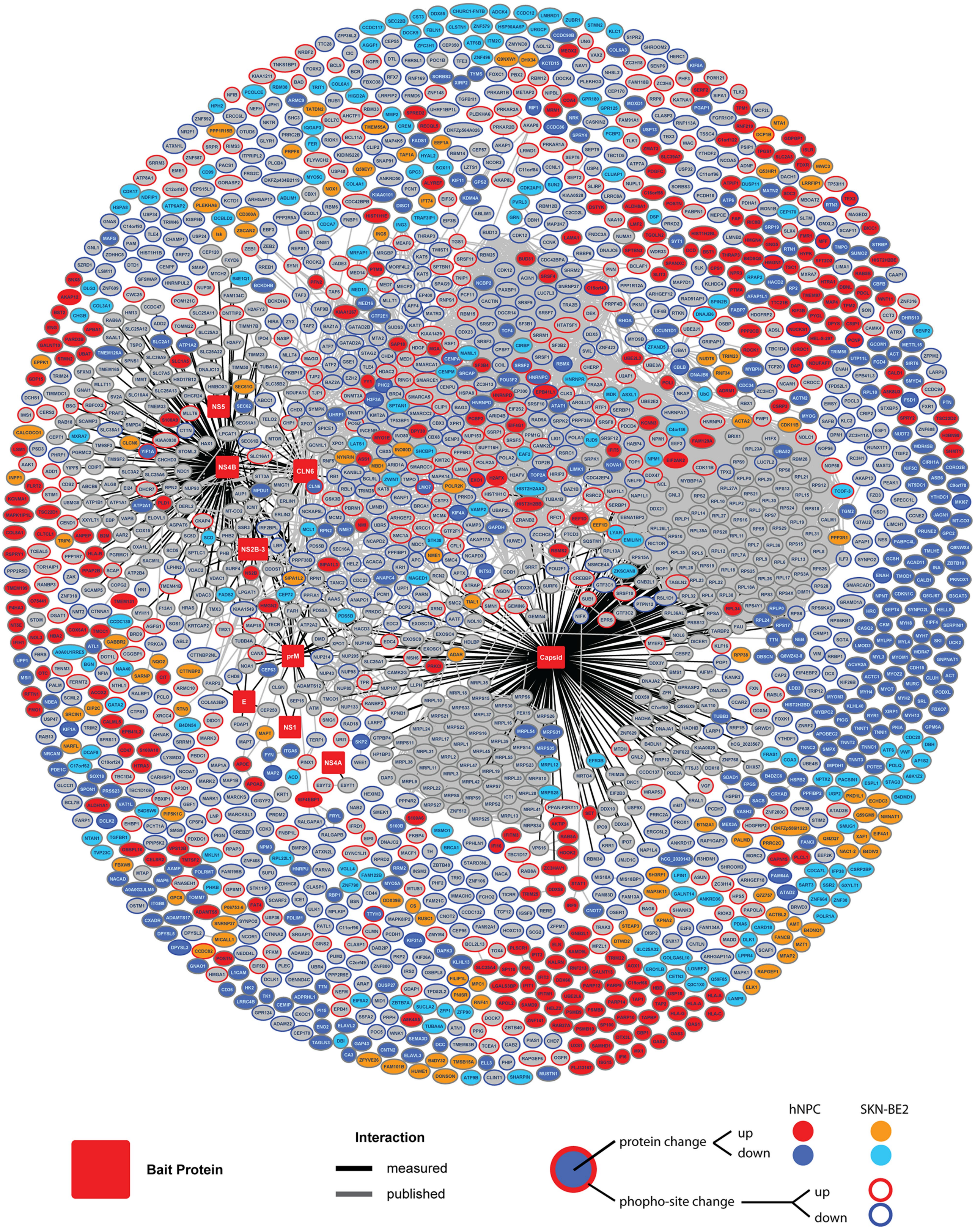


Extended Data Fig. 8 | Biological functions and pathways modulated by ZIKV at the phosphorylation level and validation by phospho-specific antibodies. a, **b**, Biological functions (**a**) and canonical pathways (**b**) modulated by ZIKV infection (IPA, $n = 4$, right-tailed Fisher's exact test, unadjusted $P \leq 0.05$. Corresponding Benjamini–Hochberg-adjusted P values are provided in Supplementary Tables 6 and 7). P, pathway; R, regulation; S, signalling. **c**–**e**, Immunoblot analysis of mock- or ZIKV-infected SK-N-BE2 with phosphorylation- or virus-specific antibodies. Where indicated, samples were treated with 12-*O*-tetradecanoylphorbol-13-acetate (TPA; 32 nM) 15 min before lysis. $n = 2$. Representative blots are shown.



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Effect of shRNA-mediated silencing of ZIKV-modulated host proteins on viral replication. a, Schematic representation of the experimental set-up used for the shRNA screen. Selected target genes were chosen among the cellular host factors specifically binding ZIKV-capsid or -NS4B (interactome; $n = 26$), significantly regulated at the phosphoproteomic level by ZIKV infection (phosphoproteome; $n = 6$) or significantly regulated at the proteomic level both by ZIKV infection and ZIKV-NS4B-expression in hNPCs (proteome; $n = 23$). Controls included non-targeting shRNA of two different pLKO-based lentivirus generations (NT1 and NT2) used as reference; and shRNAs targeting ATP6V0C and Musashi-1, which impair flavivirus pH-dependent viral entry and ZIKV replication, respectively. Two individual shRNAs per gene were selected from the MISSION TRC library (Sigma-Aldrich), and used for lentivirus production in HEK293T cells as described in the Methods. SK-N-BE2 cells were transduced with individual lentiviruses (three wells per shRNA) in three independent experiments. Three days post-transduction, cells were infected with ZIKV H/PF/2013 (MOI = 0.1) and 48 h later virus-containing supernatants were used for titration by PFU assay. The cellular viability was assessed with a resazurin-based assay that was performed in parallel. **b**, **c**, ZIKV titres and cell viability upon knockdown of selected host factors. Viral titres, as determined by PFU assay, and cell viability, as determined by resazurin assay, are expressed as a percentage of the non-targeting controls. $n = 3$ independent experiments. The box plot middle line corresponds to the median, the hinges represent the first and third quartiles, the lower (upper) whisker extends from the lower (upper) hinge to the minimum (maximum) value no further than 1.5× the interquartile range. **d**, Cell viability of SK-N-BE2 cells upon knockdown of selected host factors ($n = 4$ independent experiments, normalized mean ± s.d.; related to Fig. 3b). **e**, Validation of the silencing efficiency of selected shRNA targeting ZIKV host factors. Gene silencing was evaluated for each cellular protein via western blot detection using the antibodies specified on the right. Asterisk indicates non-specific bands, numbers on the left indicate molecular weight markers expressed in kDa. $n = 2$ independent experiments. Representative blots are shown.


NS5
NS4B
CLN6
NS2B-3
prM
Capsid
E
NS1
NS4A
Bait Protein
Interaction
measured
published
protein change
up
down
phopho-site change
up
down
hNPC
SKN-BE2

Extended Data Fig. 10 | Integration of data from orthogonal proteomic screens and protein–protein interaction databases. Integrated network of ZIKV-interacting proteins, CLN6-interacting cellular proteins, proteins changing at proteome and/or phosphoproteome levels plus measured or published interactions between them. Baits are shown as large red squares. Solid lines represent specific interactions identified by AP–MS/MS analysis, grey lines represent published protein–protein interactions from IntAct and CORUM databases. Up- or downregulation at the proteome level is marked with filled circles (hNPCs: up, red; down, dark blue; SK-N-BE2: up, orange; down, light blue). Phosphorylation changes in SK-N-BE2 are presented as red (up) or blue (down) circle borders.

LETTER

https://doi.org/10.1038/s41586-018-0448-9

Parkin and PINK1 mitigate STING-induced inflammation

Danielle A. Sliter[1], Jennifer Martinez[2], Ling Hao[1], Xi Chen[3], Nuo Sun[4], Tara D. Fischer[1], Jonathon L. Burman[1], Yan Li[5], Zhe Zhang[1], Derek P. Narendra[6], Huaibin Cai[3], Max Borsche[7], Christine Klein[7] & Richard J. Youle[1]*

Although serum from patients with Parkinson's disease contains elevated levels of numerous pro-inflammatory cytokines including IL-6, TNF, IL-1β, and IFNγ, whether inflammation contributes to or is a consequence of neuronal loss remains unknown[1]. Mutations in parkin, an E3 ubiquitin ligase, and PINK1, a ubiquitin kinase, cause early onset Parkinson's disease[2,3]. Both PINK1 and parkin function within the same biochemical pathway and remove damaged mitochondria from cells in culture and in animal models via mitophagy, a selective form of autophagy[4]. The in vivo role of mitophagy, however, is unclear, partly because mice that lack either PINK1 or parkin have no substantial Parkinson's-disease-relevant phenotypes[5–7]. Mitochondrial stress can lead to the release of damage-associated molecular patterns (DAMPs) that can activate innate immunity[8–12], suggesting that mitophagy may mitigate inflammation. Here we report a strong inflammatory phenotype in both *Prkn*$^{-/-}$ and *Pink1*$^{-/-}$ mice following exhaustive exercise and in *Prkn*$^{-/-}$;*mutator* mice, which accumulate mutations in mitochondrial DNA (mtDNA)[13,14]. Inflammation resulting from either exhaustive exercise or mtDNA mutation is completely rescued by concurrent loss of STING, a central regulator of the type I interferon response to cytosolic DNA[15,16]. The loss of dopaminergic neurons from the substantia nigra pars compacta and the motor defect observed in aged *Prkn*$^{-/-}$;*mutator* mice are also rescued by loss of STING, suggesting that inflammation facilitates this phenotype. Humans with mono- and biallelic *PRKN* mutations also display elevated cytokines. These results support a role for PINK1- and parkin-mediated mitophagy in restraining innate immunity.

If mitophagy can function to remove damaged mitochondria to reduce the release of damage-associated molecular patterns (DAMPs), *Pink1*$^{-/-}$ and *Prkn*$^{-/-}$ mice would be expected to display increased levels of inflammatory cytokines. However, we found that cytokine levels in serum of *Prkn*$^{-/-}$ and *Pink1*$^{-/-}$ mice were similar to those in wild-type mice (Extended Data Fig. 1a). To acutely stress mitochondria, 12-week-old mice were exercised until exhausted for three consecutive days. No differences in the time to exhaustion were observed among wild-type, *Prkn*$^{-/-}$ or *Pink1*$^{-/-}$ mice (Extended Data Fig. 1b). Upon mitochondrial damage, PINK1 is stabilized on the outer mitochondrial membrane where it phosphorylates ubiquitin on Ser65, forming phospho-ubiquitin (pS65-Ub), which recruits and activates parkin, inducing mitophagy[4]. Therefore, to determine whether exhaustive exercise (EE) activates PINK1 in vivo, we quantified pS65-Ub levels in heart tissue. In wild-type heart, EE induced a twofold increase in pS65-Ub relative to control sedentary (SED) mice (Fig. 1a). In SED *Pink1*$^{-/-}$ mice, pS65-Ub levels were lower than in wild-type SED mice and did not increase with EE (Fig. 1a). *Prkn*$^{-/-}$ mice also did not display increased pS65-Ub following EE (Extended Data Fig. 1c). We directly measured mitophagy in vivo in heart tissue of wild-type and *Pink1*$^{-/-}$ mt-Keima mice[17]. Consistent with pS65-Ub levels, mitophagy increased twofold following EE in wild-type heart tissue relative to SED and was significantly lower following EE in *Pink1*$^{-/-}$ mice relative to wild-type mice (Fig. 1b, c). Therefore, EE triggers PINK1 activation and mitophagy in vivo.

When mitophagy was increased in wild-type mice, there was no change in the serum concentration of IL-6, IFNβ1, IL-12(p70), IL-13, CXCL1, CCL2 or CCL4 either immediately or 24 hours after EE (Fig. 1d, e, Extended Data Fig. 1e–i). However, serum concentrations of these cytokines were markedly increased in serum of mitophagy-deficient *Prkn*$^{-/-}$ and *Pink1*$^{-/-}$ mice (Fig. 1d, e, Extended Data Fig. 1e–i). We found no differences in cytokines between SED wild-type and *Prkn*$^{-/-}$ or *Pink1*$^{-/-}$ mice, or between wild-type and *Lrrk2*$^{G2019S/G2019S}$ mice following EE (Extended Data Fig. 1d, l). Moreover, after EE, *Pink1*$^{+/-}$ mice had increased IL-6, IFNβ1, IL-12(p70), CXCL1 and CCL4, and *Prkn*$^{+/-}$ mice displayed increased IL-6 (Extended Data Fig. 2a–f). We also examined serum from human *PRKN* and *PINK1* heterozygotes (Supplementary File 1). Consistent with previous reports, levels of IL-6, IL-1β, CCL2, and CCL4 were higher in serum from patients with idiopathic Parkinson's disease than in serum from healthy controls (Fig. 2a, Extended Data Fig. 2g–i). Relative to control serum, levels of IL-6, IL-1β, CCL2, and CCL4 in unaffected *PRKN* heterozygotes were significantly higher, and were comparable to levels in patients with idiopathic Parkinson's disease or patients with biallelic *PRKN* mutations, whereas the levels of these cytokines in serum of *PINK1* heterozygotes were similar to those in control serum (Fig. 2a, Extended Data Fig. 2g–i). Therefore, loss of Parkin also leads to an increase in inflammatory cytokines in humans before, or independent of, symptomatic neurodegeneration.

IL-6 is an endogenous pyrogen. As EE in *Prkn*$^{-/-}$ and *Pink1*$^{-/-}$ mice induced an increase of about 1000-fold in serum IL-6, we measured the body temperature of these mice. One day after EE, the surface body temperatures of *Prkn*$^{-/-}$ and *Pink1*$^{-/-}$ mice, but not those of wild-type mice, were significantly increased and remained elevated for 3 to 4 days before returning to baseline (Fig. 2b), consistent with the changes in serum cytokine levels (Fig. 2c, d, Extended Data Fig. 3d–i).

Mitochondrial dysfunction can trigger inflammation via the NLRP3 inflammasome and/or the cGAS–STING pathway[8–12]. Although mitophagy can mitigate NLRP3-induced inflammation in vitro[18], IL-1β, a product of NLRP3 activation[19], was not increased in *Prkn*$^{-/-}$ or *Pink1*$^{-/-}$ EE mice (Extended Data Fig. 1j), whereas the type I interferon IFNβ1, a product of STING activation[15,16], was increased (Fig. 1e). Therefore, we crossed *Prkn*$^{-/-}$ and *Pink1*$^{-/-}$ mice with STING *goldenticket* (STING$^{gt/gt}$; *Sting* is also known as *Tmem173*) mice[20] and subjected the crossed mice to EE. Time to exhaustion and baseline cytokine levels in STING$^{gt/gt}$, *Prkn*$^{-/-}$;STING$^{gt/gt}$ and *Pink1*$^{-/-}$;STING$^{gt/gt}$ mice were similar to those of wild-type mice (Extended Data Fig. 3a–c). In stark contrast to *Prkn*$^{-/-}$ and *Pink1*$^{-/-}$ mice following EE, *Prkn*$^{-/-}$;STING$^{gt/gt}$ and *Pink1*$^{-/-}$;STING$^{gt/gt}$ mice

---

[1]Biochemistry Section, Surgical Neurology Branch, National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD, USA. [2]Immunity, Inflammation, and Disease Laboratory, National Institute of Environmental Health Sciences, National Institutes of Health, Research Triangle Park, NC, USA. [3]Transgenics Section, Laboratory of Neurogenetics, National Institute of Aging, National Institutes of Health, Bethesda, MD, USA. [4]Center for Molecular Medicine, National Heart, Lung and Blood Institute, National Institutes of Health, Bethesda, MD, USA. [5]The Protein/Peptide Sequencing Facility, National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD, USA. [6]Neurogenetics Branch, National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD, USA. [7]Institute of Neurogenetics, University of Lübeck, Lübeck, Germany. *e-mail: youler@ninds.nih.gov

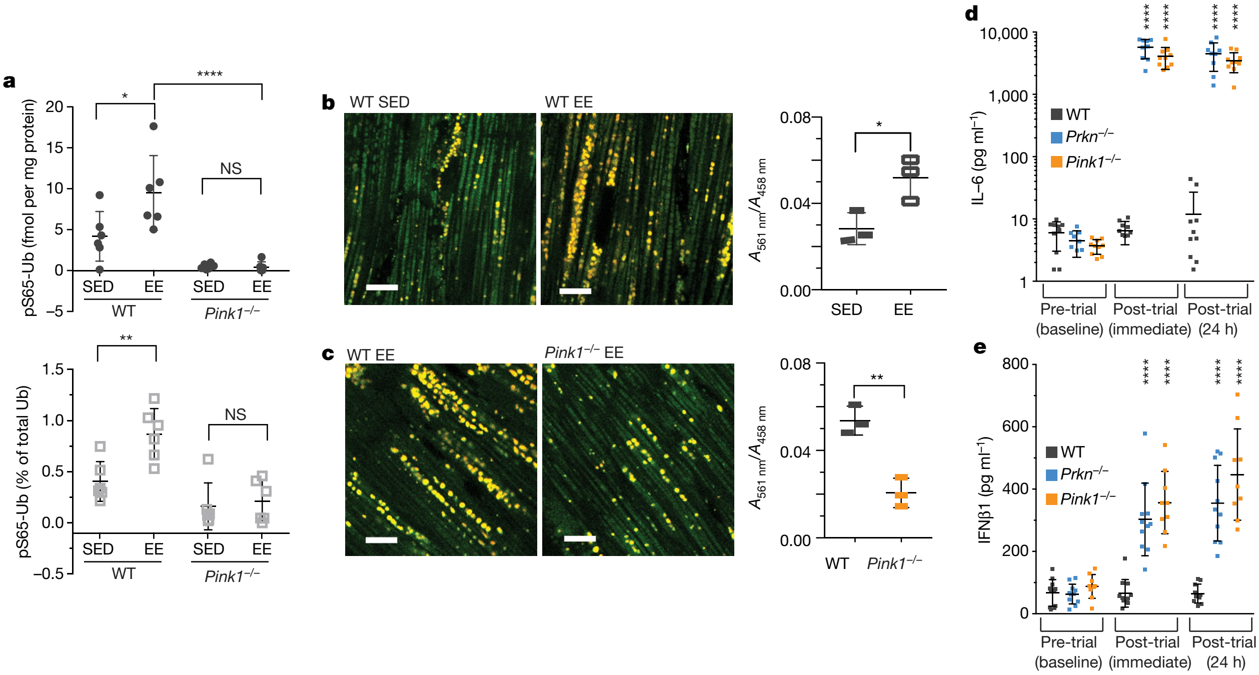


Fig. 1 | Inflammation and mitophagy in *Prkn*$^{-/-}$ and *Pink1*$^{-/-}$ mice. **a**, pS65-Ub in heart tissue, expressed as fmol per mg total protein (top), and as a percentage of total ubiquitin (bottom) ($n = 6$). WT, wild type. **b**, **c**, Left, representative images of heart tissue from SED or EE wild-type mice (**b**) or EE wild-type or *Pink1*$^{-/-}$ mice (**c**) expressing mt-Keima; right, $A_{561\,\mathrm{nm}}/A_{488\,\mathrm{nm}}$ ($n = 3$). Scale bars, 10 μm. **d**, **e**, Serum IL-6 and IFNβ1 concentrations in EE mice ($n = 10$). Data are mean ± s.d; individual data points are shown. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.005$, $^{****}P < 0.001$, NS, not significant.

showed no detectable increase in the assayed cytokines (Fig. 2c, d; Extended Data Fig. 3d–i). Consistent with this finding, surface body temperature did not increase following EE in the absence of STING (Fig. 2b).

Since STING is activated when double-stranded DNA binds cyclic GMP-AMP synthase (cGAS), which in turn generates cyclic GMP-AMP (cGAMP)[15], we examined serum DNA before and after EE. Both mtDNA copy number and the ratio of mitochondrial to nuclear DNA

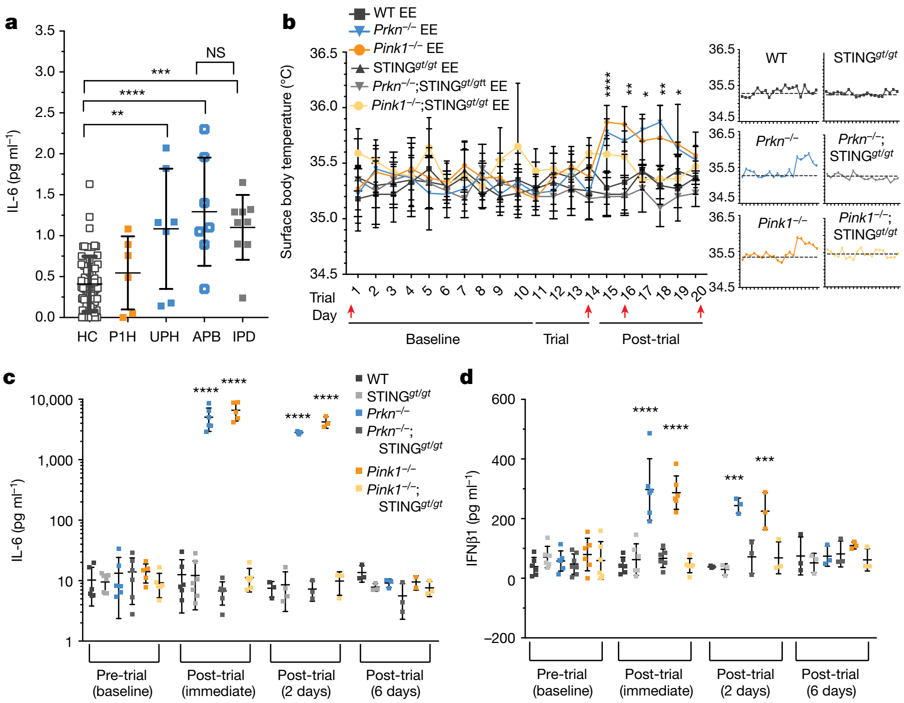


Fig. 2 | Inflammation in *Prkn*$^{-/-}$ and *Pink1*$^{-/-}$ mice following EE is completely rescued by loss of STING. **a**, Serum IL-6 concentration in humans: human control (HC) ($n = 62$), *PINK1* heterozygotes (P1H) ($n = 6$), unaffected *PRKN* heterozygotes (UPH) ($n = 7$), affected *PRKN* biallelic mutants (APB) ($n = 7$) and patients with idiopathic Parkinson's disease (IPD) ($n = 9$). **b**, Left, mean daily surface body temperature during the trial. Red arrows indicate retro-orbital serum sampling. Right, average daily temperature relative to the average baseline temperature (grey dashed line) ($n = 6$). **c**, **d**, Serum IL-6 and IFNβ1 concentrations in EE mice before and after EE. $n = 6$, baseline and post-trial immediate; $n = 3$, post-trial, 2 days; $n = 3$, post-trial, 6 days). Data are mean ± s.d. Individual data points are shown (**a**, **c**, **d**).

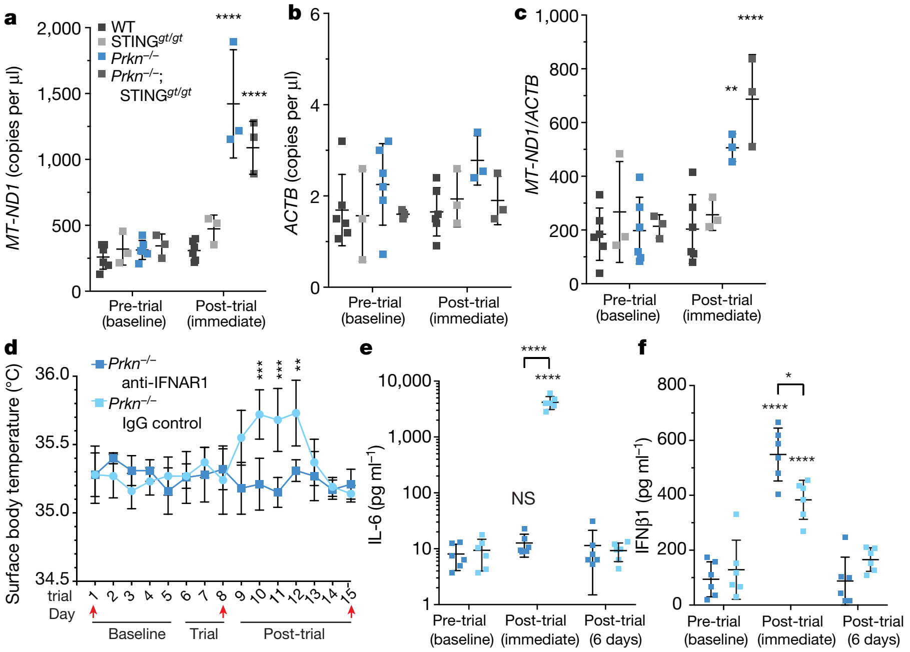


Fig. 3 | Increased circulating mtDNA in *Prkn*$^{-/-}$ mice; IFNAR1-blocking antibody treatment blocks inflammation. a, **b**, Cell-free mtDNA (*MT-ND1*, **a**) and nuclear DNA (*ACTB*, **b**) in the serum of mice with the indicated genotype ($n \geq 3$). **c**, Ratio of mtDNA to nuclear DNA in serum ($n \geq 3$). **d**, Mean daily surface body temperature during the trial ($n = 6$). Red arrows indicate retro-orbital sampling. **e**, **f**, Serum IL-6 and IFNβ1 concentrations in EE mice ($n = 6$). Data are mean ± s.d. Individual data points are shown (**a**–**c**, **e**, **f**).

increased in the serum of *Prkn*$^{-/-}$ and *Prkn*$^{-/-}$;STING$^{gt/gt}$ mice, but not in wild-type or STING$^{gt/gt}$ mice, following EE (Fig. 3a–c). Additionally, liquid chromatography–mass spectrometry analysis revealed that 2′,3′-cGAMP was markedly and comparably increased in heart tissue of *Pink1*$^{-/-}$ and *Prkn*$^{-/-}$ mice following EE, but was not detectable in wild-type or SED mice (Extended Data Fig. 4a). Treatment with a blocking antibody specific for interferon-α/β receptor subunit 1 (IFNAR1)[21] inhibited the increase in body temperature and all serum cytokines except IFNβ1 (Fig. 3d–f, Extended Data Fig. 4c–g).

To determine whether EE-induced inflammation leads to tissue damage, we measured serum creatine kinase[22]. All genotypes had similar levels of serum creatine kinase at baseline, which increased following EE in *Prkn*$^{-/-}$ and *Pink1*$^{-/-}$ mice, but not in wild-type mice (Extended Data Fig. 5a). Notably, the increase in serum creatine kinase was not rescued by loss of STING or by pretreatment with IFNAR1-blocking antibody (Extended Data Fig. 5a, b), suggesting that mitophagy, as well as being important for mitigating inflammation, may also be critical for preventing muscle damage. This reveals another conditional phenotype in *Prkn*$^{-/-}$ and *Pink1*$^{-/-}$ mice, which is potentially related to the degeneration of flight muscles observed in parkin mutant *Drosophila*[23]. Notably, parkin mutant *Drosophila* transcriptionally upregulate innate immune genes[24].

We also investigated inflammation in a chronic model of mitochondrial stress. Mice expressing a proofreading-defective mtDNA polymerase (*PolG*), called mutator, accumulate mutations in mtDNA, but do not exhibit neurodegeneration[13,14]. However, *mutator* mice deficient in parkin exhibit dopaminergic neuron loss and a movement disorder that can be rescued by treatment with levodopa[13]. Serum cytokine levels were similar among 12, 20 and 40-week-old wild-type, *mutator* and *Prkn*$^{-/-}$ mice (Fig. 4a, b, Extended Data Fig. 6a–h). However, *Prkn*$^{-/-}$;*mutator* mice exhibited higher serum levels of IL-6, IFNβ1, TNF, IL-1β, CCL2, IL-12(p70), IL-13, IL-17, CXCL1 and CCL4 at both 20 and 40 weeks of age, demonstrating considerable overlap with cytokines that are elevated in the EE-induced model of acute mitochondrial stress and human patients with Parkinson's disease (Fig. 4a–c, Extended Data Fig. 6a–h). Consistent with the EE model, loss of STING in *Prkn*$^{-/-}$;*mutator* mice fully suppressed the inflammatory phenotype (Fig. 4a, b, Extended Data Fig. 6a–h). STING activation by binding of cGAS to cytosolic double-stranded DNA (dsDNA), including mtDNA, and STING-mediated inflammation resulting from accumulation of mtDNA mutations in *mutator* mice indicate that mtDNA is the key inflammatory signal in the absence of parkin. Indeed, circulating mtDNA levels and ratios of mtDNA to nuclear DNA were higher in 40-week-old *Prkn*$^{-/-}$;*mutator* mice compared to wild-type, *mutator* or *Prkn*$^{-/-}$ mice, and loss of STING did not rescue the increase (Extended Data Fig. 7a–c). Unlike in the EE model, creatine kinase levels were not increased in aged *Prkn*$^{-/-}$;*mutator* mice (Extended Data Fig. 5c).

To determine whether STING-mediated inflammation contributes to the motor defect previously reported in *Prkn*$^{-/-}$;*mutator* mice[13], we examined locomotor activity using a pole test that is sensitive to loss of dopaminergic neurons in mouse models of Parkinson's disease[25,26]. There were no differences in the latency to descend among genotypes at 12 and 20 weeks of age (Fig. 4d, Extended Data Fig. 7d). However, 40-week-old *Prkn*$^{-/-}$;*mutator* mice displayed a markedly increased latency to descend (Fig. 4d, Supplementary Video 1). The 40-week-old *Prkn*$^{-/-}$;*mutator*;STING$^{gt/gt}$ mice displayed latency times to descend that were indistinguishable from those of wild-type and *Prkn*$^{-/-}$ mice (Fig. 4d, Supplementary Videos 2, 3). Furthermore, whereas 40-week-old *Prkn*$^{-/-}$;*mutator* mice displayed a reduction of around 20% in neurons expressing tyrosine hydroxylase (TH^{+}), concurrent loss of STING in the *Prkn*$^{-/-}$;*mutator*;STING$^{gt/gt}$ mice fully rescued TH^{+} neurons to wild-type levels (Fig. 4e, f). Even at 52 weeks of age, when *Prkn*$^{-/-}$;*mutator* mice lose about 40% of TH^{+} cells[13], *Prkn*$^{-/-}$;*mutator*;STING$^{gt/gt}$ mice displayed TH^{+} cell levels similar to those of wild-type mice (Extended Data Fig. 7e, f). Together, these data demonstrate a complete rescue of the motor defect and neurodegeneration in *Prkn*$^{-/-}$;*mutator* mice in the absence of STING.

Parkin has been linked to adaptive immunity[27], and indeed, six weeks following EE, *Prkn*$^{-/-}$ mice exhibited a marked increase in antinuclear antibodies (ANAs), whereas *Prkn*$^{-/-}$ mice treated with IFNAR1 antibody or lacking STING activity did not (Extended Data Fig. 7g). Whereas antibodies against dsDNA were not detected in mice before or after EE, they (Extended Data Fig. 7h), but not ANAs, were significantly increased in both 20 and 40-week-old *Prkn*$^{-/-}$;*mutator* mice but not in *Prkn*$^{-/-}$;*mutator*;STING$^{gt/gt}$ mice. Moreover, as in the chronic

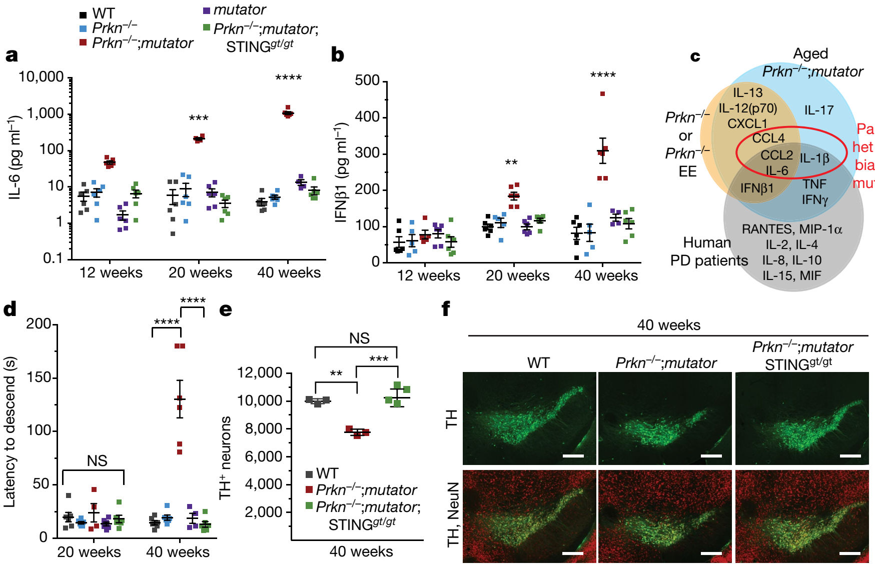


Fig. 4 | Loss of STING prevents inflammation, motor defect and neurodegeneration in the *Prkn*$^{-/-}$;*mutator* mice. a, **b**, Serum IL-6 and IFNβ1 concentrations from 12, 20 and 40-week-old mice ($n = 4$ (IL-6), $n = 6$ (IFNβ1)). **c**, Venn diagram showing serum cytokines that were elevated in mouse models and in human patients with idiopathic Parkinson's disease (PD; grey)[1]. **d**, Results of pole test ($n = 6$). **e**, Neurons expressing tyrosine hydroxylase (TH$^+$) counted by stereology in the substantia nigra of 40-week-old mice ($n = 3$ for wild type and *Prkn*$^{-/-}$;*mutator*, and $n = 4$ for *Prkn*$^{-/-}$;*mutator*;STING$^{gt/gt}$). **f**, Representative images of TH$^+$ (green) and NeuN$^+$ (total, red) neurons. Scale bars, 200 μm. Data are mean ± s.d; individual data points are shown.

Prkn$^{-/-}$;*mutator* stress model, the level of dsDNA antibodies, but not that of ANAs, was increased in the serum of patients with Parkinson's disease[28].

Here, we demonstrate that acute (EE-induced) and chronic (mtDNA mutation-induced) in vivo mitochondrial stress leads to a STING-meditated type I interferon response in mice in the absence of parkin or PINK1. Additionally, we found elevated cytokine levels in serum of asymptomatic human heterozygous *Prkn* mutation carriers. Intriguingly, intake of non-steroidal anti-inflammatory drugs is inversely associated with a later diagnosis of Parkinson's disease[29]. Therefore, we hypothesize that parkin and PINK1 prevent inflammation and neurodegeneration by clearing damaged mitochondria, thereby preventing increases in cytosolic and circulating mtDNA, suggesting a new model for how mitophagy may mitigate Parkinson's disease. Supporting our findings, a recent report demonstrated that dopaminergic neurons in the substantia nigra have basal mitophagy rates higher than those of other dopamine-producing neurons[30].

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0448-9

Received: 3 February 2018; Accepted: 23 July 2018;
Published online 22 August 2018.

1. Dzamko, N., Geczy, C. L. & Halliday, G. M. Inflammation is genetically implicated in Parkinson's disease. *Neuroscience* **302**, 89–102 (2015).
2. Kitada, T. et al. Mutations in the parkin gene cause autosomal recessive juvenile parkinsonism. *Nature* **392**, 605–608 (1998).
3. Valente, E. M. et al. PINK1 mutations are associated with sporadic early-onset parkinsonism. *Ann. Neurol.* **56**, 336–341 (2004).
4. Pickrell, A. M. & Youle, R. J. The roles of PINK1, parkin, and mitochondrial fidelity in Parkinson's Disease. *Neuron* **85**, 257–273 (2015).
5. Goldberg, M. S. et al. Parkin-deficient mice exhibit nigrostriatal deficits but not loss of dopaminergic neurons. *J. Biol. Chem.* **278**, 43628–43635 (2003).
6. Kitada, T. et al. Impaired dopamine release and synaptic plasticity in the striatum of PINK1-deficient mice. *Proc. Natl Acad. Sci. USA* **104**, 11441–11446 (2007).
7. Perez, F. A. & Palmiter, R. D. Parkin-deficient mice are not a robust model of parkinsonism. *Proc. Natl Acad. Sci. USA* **102**, 2174–2179 (2005).
8. Nakahira, K. et al. Autophagy proteins regulate innate immune responses by inhibiting the release of mitochondrial DNA mediated by the NALP3 inflammasome. *Nat. Immunol.* **12**, 222–230 (2011).
9. Zhou, R., Yazdi, A. S., Menu, P. & Tschopp, J. A role for mitochondria in NLRP3 inflammasome activation. *Nature* **469**, 221–225 (2011).
10. Rongvaux, A. et al. Apoptotic caspases prevent the induction of type I interferons by mitochondrial DNA. *Cell* **159**, 1563–1577 (2014).
11. White, M. J. et al. Apoptotic caspases suppress mtDNA-induced STING-mediated Type I IFN production. *Cell* **159**, 1549–1562 (2014).
12. West, A. P. et al. Mitochondrial DNA stress primes the antiviral innate immune response. *Nature* **520**, 553–557 (2015).
13. Pickrell, A. M. et al. Endogenous Parkin preserves dopaminergic substantia nigral neurons following mitochondrial DNA mutagenic stress. *Neuron* **87**, 371–381 (2015).
14. Trifunovic, A. et al. Premature ageing in mice expressing defective mitochondrial DNA polymerase. *Nature* **429**, 417–423 (2004).
15. Chen, Q., Sun, L. & Chen, Z. J. Regulation and function of the cGAS–STING pathway of cytosolic DNA sensing. *Nat. Immunol.* **17**, 1142–1149 (2016).
16. Ishikawa, H. & Barber, G. N. STING is an endoplasmic reticulum adaptor that facilitates innate immune signaling. *Nature* **455**, 674–678 (2008).
17. Sun, N. et al. Measuring in vivo mitophagy. *Mol. Cell* **60**, 685–696 (2015).
18. Zhong, Z. et al. NF-κB restricts inflammasome activation via elimination of damaged mitochondria. *Cell* **164**, 896–910 (2016).
19. Guo, H., Callaway, J. B. & Ting, J. P. Y. Inflammasomes: mechanism of action, role in disease, and therapeutics. *Nat. Med.* **21**, 677–687 (2015).
20. Sauer, J.-D. et al. The *N*-ethyl-*N*-nitrosourea-induced *goldenticket* mouse mutant reveals an essential function of *Sting* in the in vivo interferon response to listeria monocytogenes and cyclic dinucleotides. *Infect. Immun.* **79**, 688–694 (2011).
21. Sheehan, K. C. F. et al. Blocking monoclonal antibodies specific for mouse IFN-α/β receptor subunit 1 (IFNAR-1) from mice immunized by in vivo hydrodynamic transfection. *J. Interferon Cytokine Res.* **26**, 804–819 (2006).
22. Brancaccio, P., Lippi, G. & Maffulli, N. Biochemical markers of muscular damage. *Clin. Chem. Lab. Med.* **48**, 757–767 (2010).
23. Greene, J. C. et al. Mitochondrial pathology and apoptotic muscle degeneration in *Drosophila* parkin mutants. *Proc. Natl Acad. Sci. USA* **100**, 4078–4083 (2003).
24. Greene, J. C., Whitworth, A.J., Andrews, L.A., Parker, T. J., and Pallanck, L. J. Genetic and genomic studies of *Drosophila* parkin mutants implicate oxidative stress and innate immune responses in pathogenesis. *Hum. Mol. Gene.* **14**, 799–811 (2005).
25. Matsuura, K., Kabuto, H., Makino, H. & Ogawa, N. Pole test is a useful method for evaluating the mouse movement disorder caused by striatal dopamine depletion. *J. Neurosci. Methods* **73**, 45–48 (1997).
26. Ogawa, N., Hirose, Y., Ohara, S., Ono, T. & Watanabe, Y. A simple quantitative bradykinesia test in MPTP-treated mice. *Res. Commun. Chem. Pathol. Pharmacol.* **50**, 534–541 (1985).
27. Matheoud, D. et al. Parkinson's disease-related proteins PINK1 and Parkin repress mitochondrial antigen presentation. *Cell* **166**, 314–327 (2016).
28. Benkler, M. et al. Immunology, autoimmunity, and autoantibodies in Parkinson's disease. *Clin. Rev. Allergy Immunol.* **42**, 164–171 (2012).

29. Noyce, A. J. et al. Meta-analysis of early nonmotor features and risk factors for Parkinson disease. *Ann. Neurol.* **72**, 893–901 (2012).
30. Guzman, J. N. et al. Systemic isradipine treatment diminishes calcium-dependent mitochondrial oxidant stress. *J. Clin. Invest.* **128**, 2266–2280 (2018).

Acknowledgements We thank the animal husbandry staff at NINDS and the staff of the Murine Phenotyping Core Facility at NHLBI; the Clinical Pathology Group in the Cellular & Molecular Pathology Branch at NIEHS for serum creatine kinase; K. Gerrish and B. Elgart from the NIEHS Molecular Genomics Core for serum mitochondrial and nuclear DNA isolation and quantification; J. Vargas and S. Humble for experimental assistance and T. Finkel, J. Kowalak, A. Oberst and M. Ward for helpful suggestions. This work was supported by the NINDS Intramural Research Program (R.J.Y.), NIH Intramural Research Program 1ZIAES10328601 (J.M.), the NIA Intramural Research Program (H.C.) the DFG FOR2488; P2 and a Pilot Grant from the Excellence Cluster Inflammation at Interfaces (C.K.), and by a career development award from the Hermann and Lilly Schilling Foundation (C.K.).

Reviewer information *Nature* thanks Z. Chen, I. Dikic and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions The project was conceived by D.A.S., J.M. and R.J.Y. Mouse experiments including EE, sample collection and behavioural studies were conducted by D.A.S. Serum analysis was performed by J.M. Mass spectrometry was performed by L.H. and Y.L. Neuron counting was performed by X.C. and H.C. mt-Keima analysis was performed by N.S. LRRK2 mutant mice were exercised by T.D.F. J.L.B. and Z.Z. provided technical assistance. Human patient samples were provided by D.P.N., M.B. and C.K. Writing and editing by D.A.S., J.M., C.K. and R.J.Y. Funding was provided to J.M., H.C, D.P.N, C.K. and R.J.Y.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0448-9.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0448-9.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to R.J.Y.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were blinded to allocation during certain experiments and outcome assessment as specified in the methods.

Mouse strains. All mice were housed in pathogen-free facilities under 12-h light dark cycles with access to food and water ad libitum. The STING *goldenticket* (STING$^{gt/gt}$) mouse was obtained from Jackson Laboratories (C57BL/6J-Tmem173gt/J). It was generated by a chemically-induced mutation causing a T to A transversion which resulted in a point mutation at L199[19]. The point mutation triggers protein degradation and consequently, STING$^{gt/gt}$ mice do not express STING or produce type I interferons in response to cyclic di-nucleotides[19]. Wild-type (C57BL6/J), *Prkn*$^{-/-}$ (B6.129S4-Park2$^{\text{tm1Shn/J}}$), *Pink1*$^{-/-}$ (B6.129S4-Pink1$^{\text{tm1Shn/J}}$) and *mutator* (B6.129S7 (Cg)-Polg$^{\text{tm1Tprol/J}}$) mice were obtained from Jackson Laboratories. *Prkn*$^{-/-}$ and *Pink1*$^{-/-}$ mice were crossed with STING$^{gt/gt}$ mice to generate *Prkn*$^{-/-}$;STING$^{gt/gt}$ or *Pink1*$^{-/-}$;STING$^{gt/gt}$ mice. *Prkn*$^{-/-}$;*mutator* mice were generated as described[13] and crossed with *Prkn*$^{-/-}$;STING$^{gt/gt}$ mice to generate *Prkn*$^{-/-}$;*mutator*;STING$^{gt/gt}$ mice. Owing to increased mtDNA mutational load and reduced fertility all *mutator* matings were performed with heterozygous mice. *Lrrk2*$^{G2019S/G2019S}$ mice were obtained from Novartis by H. Cai (NHLBI) and generated as described[31]. mt-Keima mice on the FVB/NJ background were obtained from N. Sun, NHBLI and generated as described[16]. *Pink1*$^{-/-}$ mice were crossed with mt-Keima mice to generate *Pink1*$^{-/-}$ mice expressing mt-Keima. Genotypes were confirmed by PCR[5,6,19,32]. All mice, except mt-Keima mice, have the nuclear background of C57BL/6J. All animal studies were carried out as approved by the Animal Care and Use Committee of the National Institute for Neurological Disorders and Stroke or the National Heart, Lung and Blood Institute.

Exhaustive exercise. To induce mitochondrial stress, we subjected mice to forced exhaustive treadmill running on a Columbus Eco 3/6 (Columbus Instruments). Ten-week-old mice were randomly assigned into sedentary (SED) and exhaustive exercise (EE) groups. All groups were counterbalanced by gender. At 11 weeks old, EE mice were familiarized with the treadmill using three days of low intensity running. On day 1 of the familiarization protocol, mice placed on the treadmill were given 5 min to acclimate before the treadmill belt was engaged. The belt speed was initially set to 8m/min then increased to 10m/min, then 12m/min for 5 min each, all at an incline of 10%. On day 2, mice were given 2 min to acclimate to the treadmill then exercised at 8m/min, 10m/min, 12m/min and 15m/min for 5 min each. On day 3, mice were placed on the treadmill and immediately exercised at 10m/min for 5 min, then 12m/min, 15m/min and 18m/min for 5 min each. After familiarization, mice were allowed two days to recover before beginning the EE protocol. For three consecutive days, mice were subjected to the following protocol: At a 10% incline, the initial speed was set at 10m/min for 8 min, then increased to 15m/min for 5 min, and then increased by 1.8m/minute every three minutes until reaching a speed of 22.4m/minute. Mice were exercised for 10 min at 22.4m/min, then speed was then increased by 1m/min every 5 min until exhaustion. Exhaustion was determined when the mice failed to re-engage all four paws with the treadmill belt despite negative stimulus. For each EE experiment, the number of mice were listed per experimental group in the figure legends and individual data points were graphed. The experimenter was blinded to the genotype of the mouse being exercised.

Blood collection and analysis. Blood was drawn from the retro-orbital sinus of SED and EE mice anesthetized with isoflourane using heparinized capillary tubes. Baseline blood draws were performed on 10-week-old mice before exercise. Post-trial immediate blood draws were collected within 10 min of completion on the third day of the EE protocol. Final blood draws were performed at 24 h, 2 days, 6 days or 6 weeks after completion of the EE protocol. The same mouse was sampled for each time point (baseline, post-trial immediate and post-trial final). Serum was analysed in a blinded manner for cytokines using Bio-plex Pro Mouse Cytokine Standard 23-Plex, Group 1 (Bio-Rad) and Legend Max Mouse IFN−β ELISA kit (BioLegend) and Mouse Anti-Nuclear Antigens (ANA/ENA) Ig's (total (A+G+M)) ELISA Kit (Alpha Diagnostics). Creatine kinase (CK) was measured using a modified International Federation of Clinical Chemistry (IFCC) method[33]. Briefly, CK reversibly catalyses the transfer of a phosphate group from creatine phosphate to ADP, resulting in creatine and ATP. ATP is used to produce glucose-6-phosphate and ADP from glucose. The glucose-6-phosphate is oxidized with simultaneous reduction of the coenzyme NADP to give NADPH and 6-phosphogluconate. Therefore, the activity of CK is measured as U/ml directly proportional to the formation of NADPH. Reactions were carried as previously described, and $A_{340\text{ nm}}/A_{600\text{ nm}}$ was read on an Olympus AU400e clinical analyser (Beckman Coulter). Circulating cell-free mtDNA and nuclear DNA was isolated from serum using the Maxwell RSC ccfDNA Plasma Kit (Promega) using the Maxwell RSC Instrument (Promega), then quantified using digital droplet PCR on a Maxwell as described[34]. The assay used to detect *MT-ND1* was dMmuCNS343824284 (Bio-rad), and the assay used to detect *ACTB* was dMmuCNS292036842 (Bio-Rad).

Pole test. The pole test was performed as previously described[13,25] to test the motor coordination of mice beginning at 12 weeks old, and then repeated at 20 and 40 weeks old. Mice were placed facing upward at the top of a vertical pole and given 180 s to change orientation and descend the pole. Failure to descend or falling from the pole was assigned 180 s. Each mouse performed the trial three times and the average latency to descend was determined. One day following the pole test, the same mouse also had blood drawn for cytokine analysis via the retro-orbital sinus, resulting in three pole tests and three blood draws per mouse, at 12, 20 and 40 weeks of age.

Quantification of TH^{+} neurons by stereology. According to the mouse brain in stereotaxic coordinates, a series of coronal sections across the midbrain (40 μm per section, every third section from bregma −2.54 mm to −4.16 mm) were chosen and processed for tyrosine hydroxylase (TH) (ImmunoStar, Cat# 22941) and NeuN (Abcam, Cat# ab104225) staining, and visualized using a laser scanning confocal microscope (LSM 880; Zeiss). We examined 11–13 sections per brain. The images were captured as a projected layer at 20 μm (pinhole, 10 μm; interval, 5 μm, 3–4 layers) under 10× magnification. The number of TH^{+} neurons was assessed using the fractionator function of Stereo Investigator 11 (MicroBrightField). The sampling scheme was designed to have coefficient of error (CE) of less than or equal to 5% in order to get reliable results. To achieve suitable CE, normally 12 serial sections, with a total of 140 counting frames were assessed. The final parameters of these studies were as follows: grid size, 250 × 250 μm; frame size, 200 × 200 μm. Three or four mice were used per genotype. Counters were blinded to the genotypes of the samples.

Body temperature sampling. Surface body temperature measurements were performed on SED and EE mice using the Exergen Temporal Scanner (TAT-5000, Exergen) beginning the day after the 10-week baseline blood draw and continuing daily until mice were euthanized either 2 days or 6 days after the exhaustive running protocol was complete. On days when mice were run or bled, temperatures were collected before exercise or blood draws. Mice were restrained by hand and the scanner was drawn across the abdomen from the left shoulder to the right hip. An average of three consecutive scans were recorded and the mean was calculated. For each mouse, baseline surface temperature was determined by collecting temperature for 11 days beginning the day after the baseline blood draw until beginning of the exercise trial. The average baseline temperatures depicted in the small graph insets in Fig. 2b were determined by averaging all baseline surface temperatures for each individual mouse (11 days). Surface body temperature sampling was routinely performed between 2 and 3 pm.

Anti-IFNAR1/2 blocking antibody treatment. *Prkn*$^{-/-}$ mice were randomly assigned to control and treatment groups. Treatment began following the familiarization protocol one day before starting EE via intraperitoneal injection with 2.5 mg purified functional grade anti-IFNAR1 antibody (MAR1-5E3, catalogue number I-403, Leinco Technologies) or 2.5 mg purified functional grade anti-human interferon gamma receptor alpha as an IgG control[20] (GIR 208, catalogue number I-443, Leinco Technologies). This dose of anti-IFNAR1 has a half life of 5 days[20]. Mice received only one treatment. The experimenter was blinded to the drug treatment. Three mice were treated with IgG control or anti-IFNAR1 in two separate trials for a total of $n = 6$ animals.

Quantification of pS65 ubiquitin and cGAMP in heart tissue. Ubiquitin and pS65 ubiquitin measurements were conducted as described[13] with several modifications. In brief, EE and SED mice were perfused with phosphate-buffered saline, hearts were dissected and snap frozen in liquid nitrogen. For protein extraction, heart tissue was homogenized in an ice-cold lysis buffer (50 mM Tris/HCl pH 7.5, 150 mM NaCl, 0.1% SDS, 1% Triton, 1 mM DTT, 50 μM PR-169, 5 mM *o*-phenanthroline, protease inhibitor cocktail (Roche), and phosphatase inhibitor cocktail (Roche). For ubiquitin enrichment, 4 mg of total protein from each sample was incubated overnight with Tandem Ubiquitin Binding Entities (TUBE, LifeSensors), washed four times with lysis buffer and another four times with 50 mM ammonium bicarbonate, eluted with 6 M guanidine HCl, precipitated with TCA, spiked with 200 fmol of UB-AQUA reference peptides (JPT Peptide Technologies), digested with Trypsin/Lys-C mix (Promega), and purified by HLB solid phase extraction column (Oasis). Purified peptides were dried down and redissolved in 0.1% formic acid in 2% acetonitrile. Targeted peptide analysis was performed on an UltiMate 3000 nanoLC system coupled with a Q Exactive HF mass spectrometer (Thermo). A 60 min liquid chromatography gradient was established on an EASY-Spray C18 column (Thermo, 2 μm, 100 Å, 75 μm × 250 mm) with a flow rate of 0.3 μl/min. Buffer A was 0.1% formic acid in 2% acetonitrile, and buffer B was 0.1% formic acid and 5% DMSO in 95% acetonitrile. Targeted MS/MS scans were acquired with an isolation window of 1.4 *m/z*, higher-energy collisional dissociation (HCD) with a normalized collision energy of 25%, a resolution of 15 K, an automatic gain control of 5×10^{5}, and a maximum injection time of 50 ms. Raw data were uploaded onto Skyline 3.7[35] for data analysis and the results were exported to an Excel spread sheet for subsequent statistical analysis[13,36]. To quantify cGAMP, cGAMP was extracted and enriched from mouse heart samples following the steps described[37]. Targeted

LC–MS experiment was conducted on a Dionex Ultimate 3000 nano liquid chromatography system coupled with a Fusion Lumos Mass Spectrometer. cGAMP was separated with a 40 min gradient on an EASY-Spray C18 column (Thermo, 2 μm, 100 Å, 75 μm × 250 mm) with a flow rate of 0.3 μl/min. Buffer A was 0.1% formic acid in 2% acetonitrile, and buffer B was 0.1% formic acid and 5% DMSO in 95% acetonitrile. Target *m*/*z* was 675.1080 with an isolation window of 2 Da, a resolution of 7.5K, an automatic gain control of 2×10^5, and a maximum injection time of 32 ms. Raw data were acquired and analysed in Xcalibur software.

Assessing cardiac mitophagy using the mt-Keima transgenic mice. Cardiac mitophagy was measured using confocal microscopy as previously described[17]. In brief, for the hearts, tissues were rapidly dissected out, rinsed with cold PBS, and then further processed into small pieces (approximately 100 mm^3). These sections were placed onto a 35-mm coverglass #1.5 bottom micro-well dish (MatTek) and analysed immediately. Fluorescence of mt-Keima was imaged in two channels via two sequential excitations (458 nm, green; 561 nm, red) and using a 570 to 695-nm emission range with a Zeiss LSM 780 confocal microscope (Carl Zeiss MicroImaging) equipped with a Plan-Apochromat 20× /0.8 NA, 40× /1.4 NA, and 63× /1.40 NA oil immersion objective lens. Laser power was set at the lowest output that would allow clear visualization of the mt-Keima signal. Imaging settings were maintained with the same parameters for comparison between different experimental conditions. Representative confocal images were processed using Imaris software by contrast linear stretch only. The average of four images from each tissue sample was taken, and calculation of mitophagy based on mt-Keima signal was performed employing the original images using Zeiss ZEN software on a pixel-by-pixel basis as previously described[38].

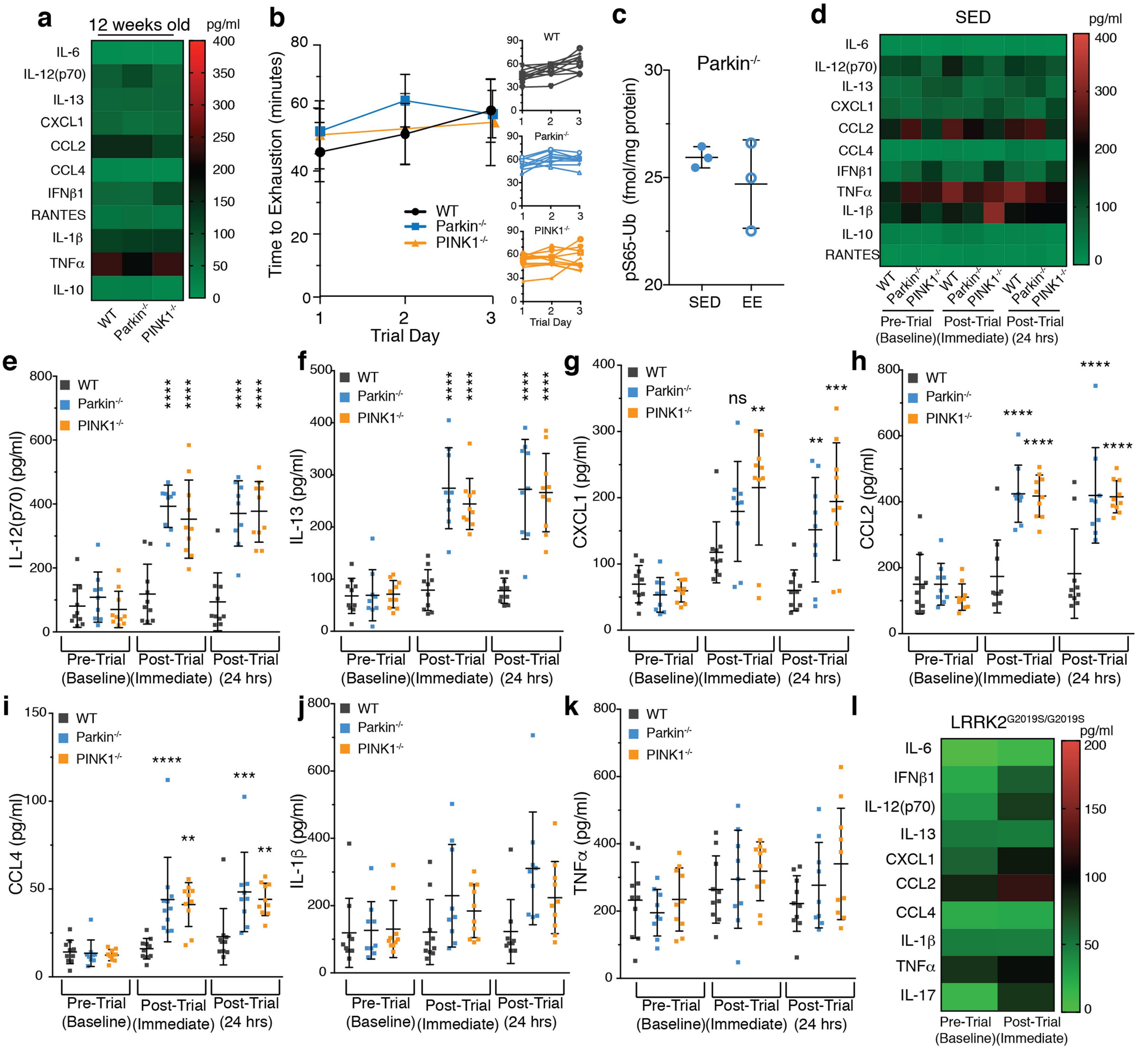
Recruitment and characterization of probands and serum sampling. All probands were recruited at the Institute of Neurogenetics and Department of Neurology of the University of Luebeck, a tertiary referral centre for movement disorders, or the National Institutes of Health (NIH) Clinical Center. Samples from the University of Lubeck were collected within the EPIPARK cohort[39], now part of the larger ProtectMove cohort which is enriched for patients with idiopathic Parkinson's disease recruited at the centre's outpatient clinics and containing a nested monogenic Parkinson's disease/asymptomatic mutation carrier cohort. The local IRB granted approval of the Parkinson's disease cohort and genetic studies. Samples from the NIH Clinical Center were collected under a protocol approved by the NIH Combined Neuroscience (CNS) Institutional Review Board. All patients and controls gave written informed consent and were neurologically examined by a movement disorder specialist. If applicable, a clinical diagnosis of Parkinson's disease was established according to the UK Brain Bank Criteria. Clinical characterization and ancillary tests included the UPDRS, Hoehn and Yahr, and MoCA score, as well as olfactory testing using the Brief Smell Identification Test. A blood sample was obtained for extraction of DNA and preparation of serum. Genetic testing was performed with the Global Screening Array containing custom content for known Parkinson's disease genes and risk factors (Illumina), Sanger sequencing of Parkinson's disease-related genes, and multiplex ligation-dependent probe amplification (MLPA) to detect gene dosage changes. Serum was analysed in a blinded manner for cytokines using Bio-Plex Pro Human Cytokine 17-plex Assay (Biorad).

Statistical analysis. Statistical analysis for multiple comparisons were performed in Prism 7 software using a two-way ANOVA on non-matched samples. All comparisons were to the same genotype, either before EE (pre-trial baseline) or to 12-week old animals shown in Fig. 4, unless otherwise noted with brackets. Human cytokines comparisons (Fig. 2a) were relative to controls unless noted by brackets. Body temperature comparisons were to the average temperature of wild-type mice measured on the same day.

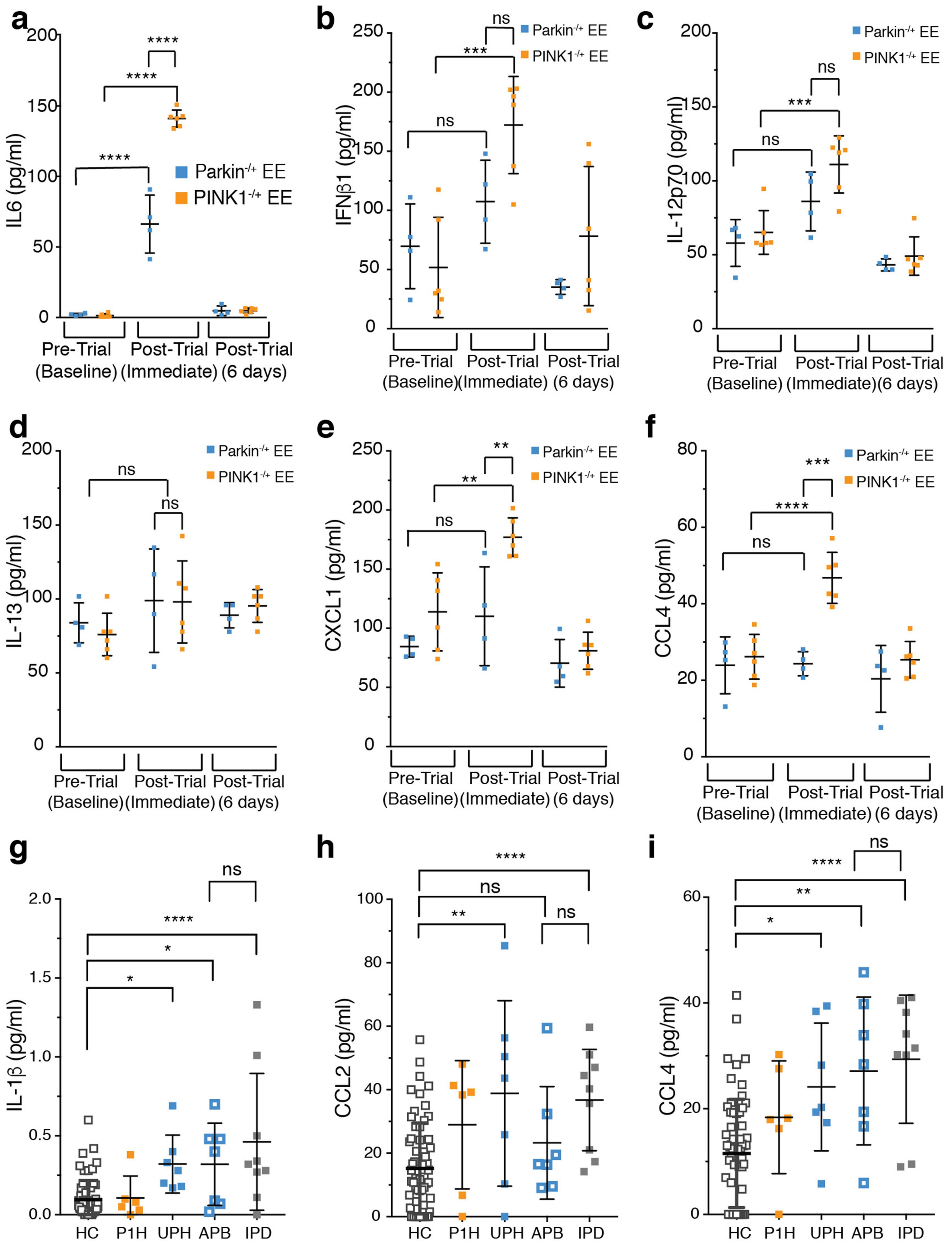
Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability. LC-MS data that support the findings of this study are freely available to the public on Open Science Framework (https://osf.io/g5zfk/?view_only=2dc5b53dcf1747379ed1ea798eb705f5) . All other relevant data are available in the manuscript.

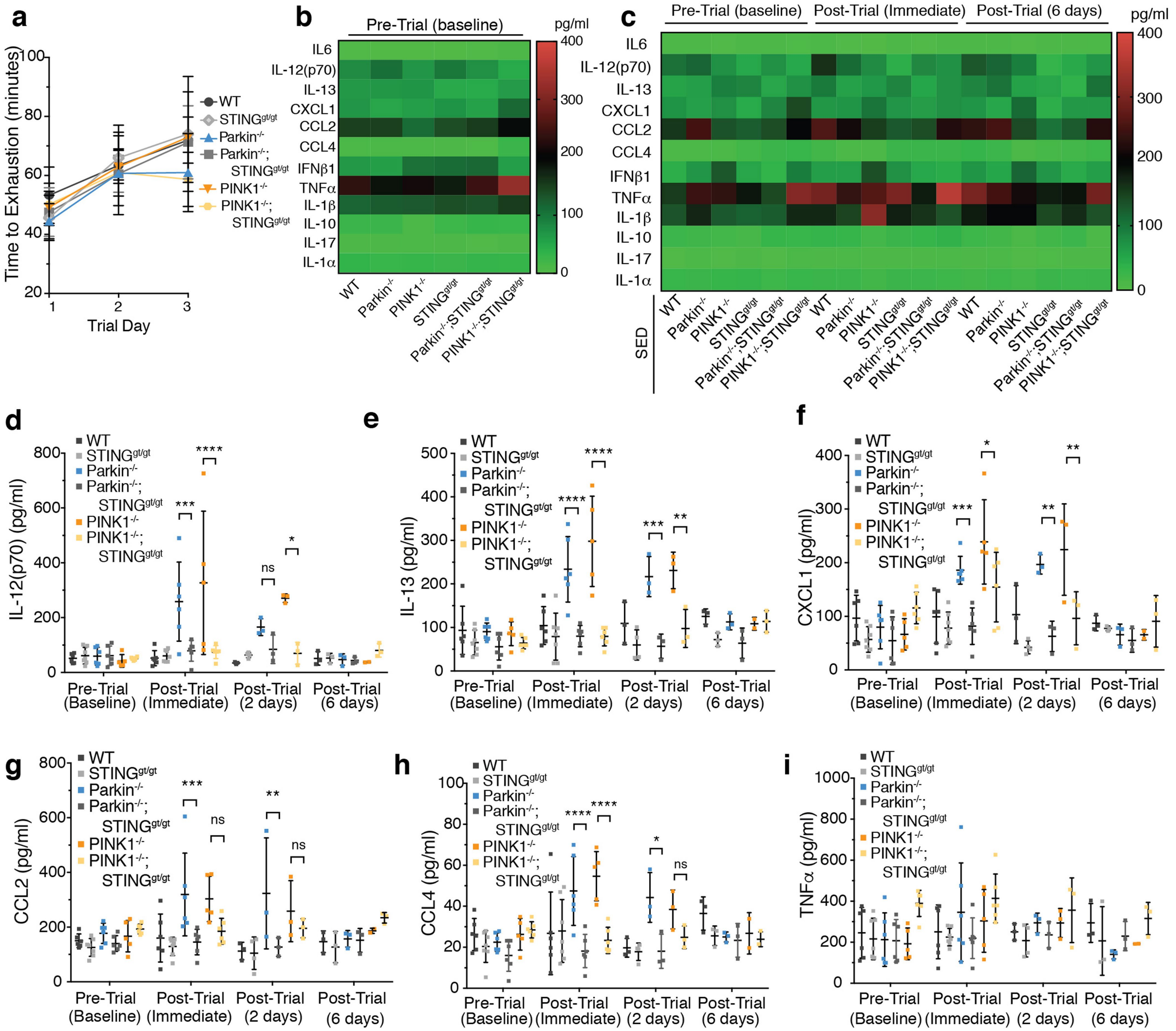
31. Herzig, M. C. et al. LRRK2 protein levels are determined by kinase function and are crucial for kidney and lung homeostasis in mice. *Hum. Mol. Genet.* **20**, 4209–4223 (2011).
32. Kujoth, G. C. et al. Mitochondrial DNA mutations, oxidative stress and apoptosis in mammalian aging. *Science* **309**, 481–484 (2005).
33. Horder, M. et al. International Federation of Clinical Chemistry, Scientific Division Committee on Enzymes: approved recommendation on IFCC methods for the measurement of catalytic concentration of enzymes. Part 7. IFCC method for creatine kinase (ATP: creatine *N*-phosphotransferase, EC 2.7.3.2). *Eur. J. Clin. Chem. Clin. Biochem*. **29**, 435–456 (1991).
34. Ye, W. et al. Accurate quantitation of circulating cell-free mitochondrial DNA in plasma by droplet digital PCR. *Anal. Bioanal. Chem*. **409**, 2727–2735 (2017).
35. MacLean, B. et al. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **26**, 966–968 (2010).
36. Phu, L. et al. Improved quantitative mass spectrometry methods for characterizing complex ubiquitin signals. *Mol. Cell. Proteomics* **10**, M110.003756 (2011).
37. Gao, D. et al. Activation of cyclic GMP-AMP synthase by self-DNA causes autoimmune diseases. *Proc. Natl Acad. Sci. USA* **112**, E5699–E5705 (2015).
38. Sun, N. et al. A fluorescence-based imaging method to measure in vitro and in vivo mitophagy using mt-Keima. *Nat. Protoc*. **12**, 1576 (2017).
39. Kasten, M., *et. al.* Cohort profile: a population based cohort of study non-motor symptoms in parkinsonism (EPIPARK). *Int. J. Epidemiol*. **42**, 128–128 (2013).



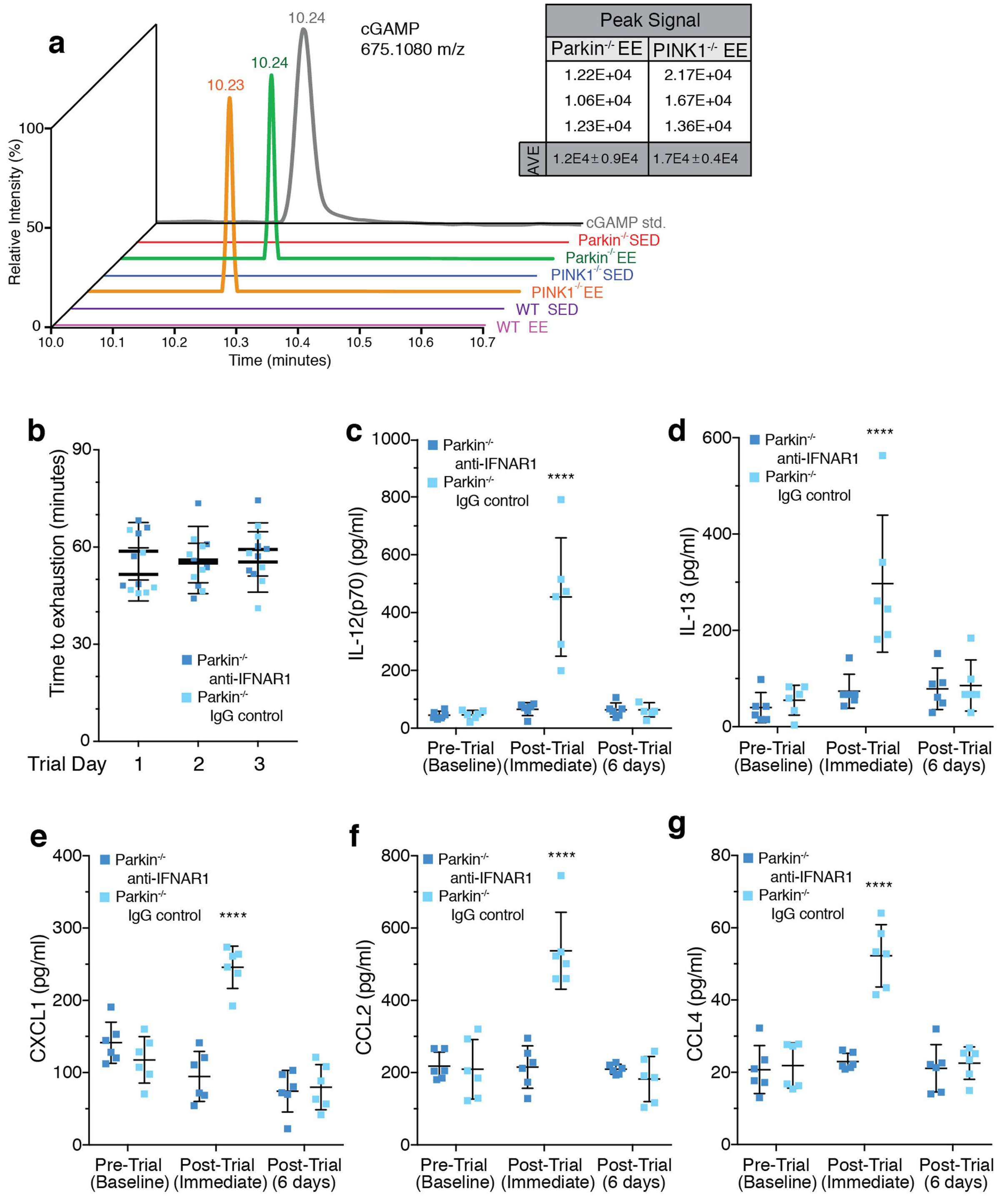
Extended Data Fig. 1 | Inflammation in *Prkn*$^{-/-}$ and *Pink1*$^{-/-}$ mice. **a**, **d**, Heat maps depicting average cytokine concentration in serum from mice ($n = 10$). **b**, Average time to exhaustion on each trial day. Small graphs show individual run times ($n = 10$). **c**, pS65-ubiquitin in *Prkn*$^{-/-}$ heart tissue expressed as fmol per mg total protein ($n = 3$). **e**–**k**, Serum cytokine concentrations from EE mice are plotted as mean ± s.d. ($n = 10$). **l**, Heat map depicting serum cytokine levels of *Lrrk2*$^{G2019S/G2019S}$ ($n = 4$). Using *t*-tests, no differences in cytokine concentrations were found between pre-trial (baseline) and post-trial (immediate). ns, not significant.



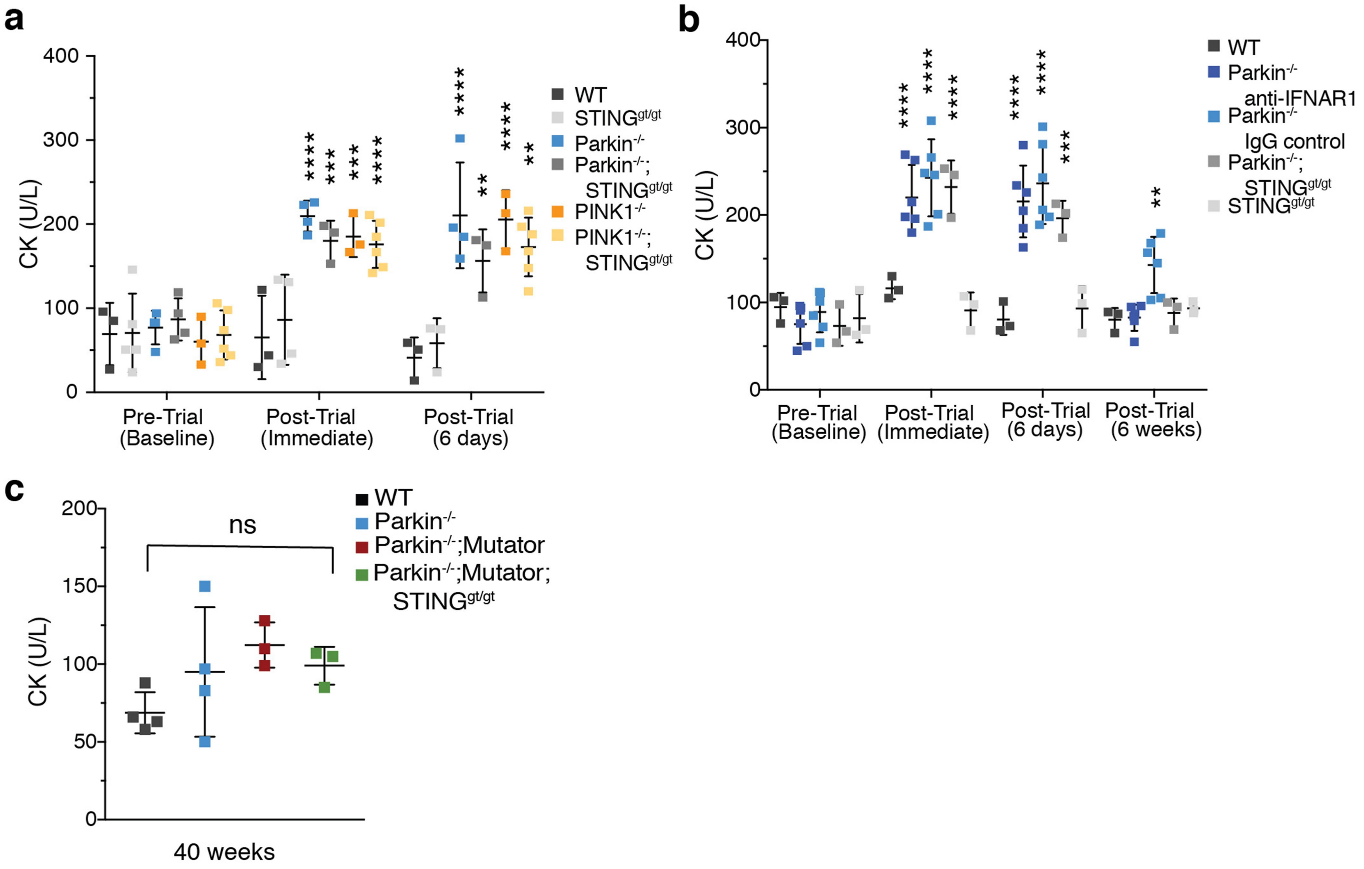
Extended Data Fig. 2 | Inflammatory cytokines are increased in mice and humans with heterozygous loss of parkin or PINK1. a–f, Serum cytokine concentrations from *Prkn*$^{-/+}$ ($n = 4$) and *Pink1*$^{-/+}$ ($n = 6$) EE mice. **g–i**, Serum cytokine concentrations from human control (HC) ($n = 62$), *PINK1* heterozygotes (P1H) ($n = 6$), unaffected *PRKN* heterozygotes (UPH) ($n = 7$), affected *PRKN* biallelic mutants (APB) ($n = 7$) and patients with idiopathic Parkinson's disease (IPD) ($n = 9$). Data are mean ± s.d.



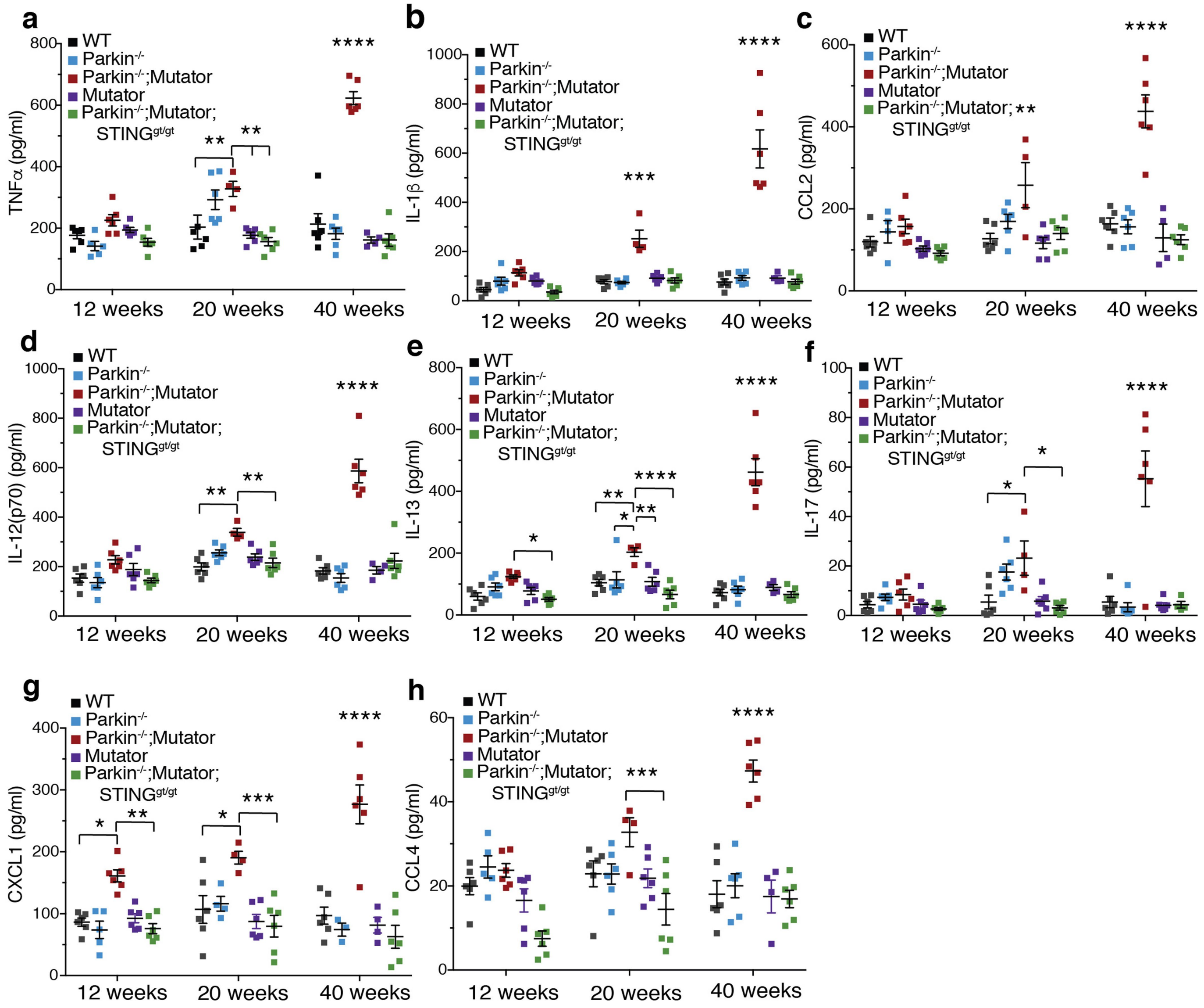
Extended Data Fig. 3 | Loss of STING prevents increased cytokine levels in *Prkn*$^{-/-}$ and *Pink1*$^{-/-}$ mice following EE. **a**, Average time to exhaustion on each trial day ($n = 6$). **b**, **c**, Heat map depicting the average baseline serum cytokine concentration for EE mice ($n = 6$) and the average serum cytokines from SED mice ($n = 6$). **d**–**i**, Serum cytokine concentrations from mice are plotted as mean ± s.d. ($n = 6$ baseline, post-trial immediate; $n = 3$, post-trial 2 days, post-trial 6 days).



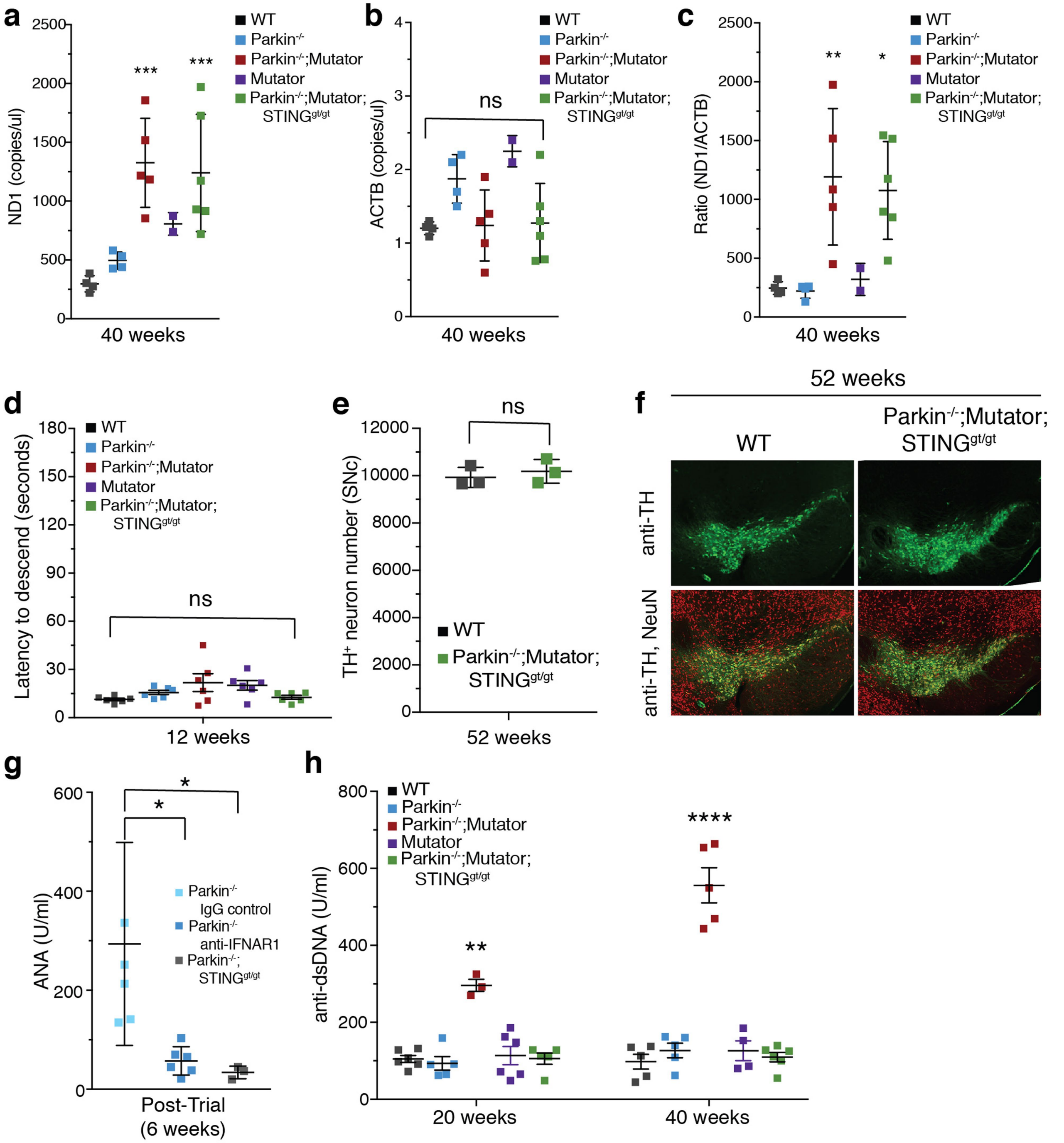
Extended Data Fig. 4 | cGAMP is increased in *Prkn*$^{-/-}$ and *Pink1*$^{-/-}$ EE heart tissue and inflammation is inhibited by anti-IFNAR1 treatment. **a**, Representative plot of signal intensity for cGAMP measured in heart tissue. cGAMP was not detected in wild-type EE or in SED mice. Average signal intensity for $n = 3$ samples is shown in the inset. **b**, Average time to exhaustion on each trial day ($n = 6$). **c**–**g**, Serum cytokine concentrations from EE *Prkn*$^{-/-}$ mice treated with anti-IFNAR1 antibody or IgG control ($n = 6$). Data are mean ± s.d.



Extended Data Fig. 5 | Elevated serum creatine kinase following EE in *Prkn*$^{-/-}$ and *Pink1*$^{-/-}$ mice is not rescued by inhibition of inflammation and is not increased by chronic mitochondrial dysfunction. a–c, Serum creatine kinase (CK) levels ($n < 3$). Data are mean ± s.d.



Extended Data Fig. 6 | Inflammation in aged *Prkn*$^{-/-}$;*mutator* mice is rescued by loss of STING. a–h, Serum cytokine concentrations from 12, 20 and 40-week-old mice ($n = 4, 6$). Data are mean ± s.d.



Extended Data Fig. 7 | STING mediates inflammation under chronic mitochondrial stress. **a**, **b**, Copy number per µl of cell-free mtDNA (ND1) or nuclear DNA (ACTB) in serum ($n < 3$). **c**, Ratio of mtDNA to nuclear DNA. ($n < 3$). **d**, The time required for 12-week-old mice to descend the pole ($n = 6$). **e**, TH^{+}-neurons counted by stereology in the substantia nigra (SNc) of 52-week-old mice ($n = 3$). **f**, Representative images of TH^{+} neurons (green) and total neurons (NeuN, red). **g**, Serum levels of antinuclear antibodies (ANA) 6 weeks post-EE. dsDNA antibodies were not detected and ANAs were not detected at baseline or immediately post-EE ($n = 4, 6$). **h**, Serum levels of anti-dsDNA antibodies ($n = 4, 6$). ANA antibodies were not detected. Data are mean ± s.d. ($n = 6$).

LETTER

https://doi.org/10.1038/s41586-018-0475-6

GAPDH inhibits intracellular pathways during starvation for cellular energy homeostasis

Jia-Shu Yang[1], Jia-Wei Hsu[1], Seung-Yeol Park[1], Jian Li[1], William M. Oldham[2], Galina V. Beznoussenko[3], Alexander A. Mironov[3], Joseph Loscalzo[4] & Victor W. Hsu[1]*

Starvation poses a fundamental challenge to cell survival. Whereas the role of autophagy in promoting energy homeostasis in this setting has been extensively characterized[1], other mechanisms are less well understood. Here we reveal that glyceraldehyde 3-phosphate dehydrogenase (GAPDH) inhibits coat protein I (COPI) transport by targeting a GTPase-activating protein (GAP) towards ADP-ribosylation factor 1 (ARF1) to suppress COPI vesicle fission. GAPDH inhibits multiple other transport pathways, also by targeting ARF GAPs. Further characterization suggests that this broad inhibition is activated by the cell during starvation to reduce energy consumption. These findings reveal a remarkable level of coordination among the intracellular transport pathways that underlies a critical mechanism of cellular energy homeostasis.

The activation of ADP-ribosylation factor (ARF) small GTPases initiates vesicular transport by recruiting coat proteins to intracellular membrane compartments for vesicle formation[2]. In the case of vesicle formation by the COPI complex, the GAP that de-activates ARF1—known as ARFGAP1—is also an ARF effector, acting as a coat component[3,4]. To gain further insight into how ARFGAP1 acts in COPI transport, we sought to identify new interacting proteins. By incubating ARFGAP1 with cytosol in a pull-down experiment, we identified GAPDH as one such protein (Extended Data Table 1). Although it is well known for acting in glycolysis, GAPDH is also known to have multiple non-glycolytic roles[5]. Therefore, to investigate whether it has a role in COPI transport, we initially performed a COPI transport assay as described[6–8]. In cells treated with small interfering RNA (siRNA) against *GAPDH*, we observed enhanced COPI transport (Fig. 1a, Extended Data Fig. 1a). Targeting specificity was confirmed by a rescue experiment (Fig. 1a, Extended Data Fig. 1b), and by treatment with an unrelated siRNA sequence (Extended Data Fig. 1c). The siRNA treatments were limited to two days to maintain cell viability (Extended Data Fig. 1d). Consistent with the effect of reducing GAPDH level, GAPDH overexpression had the opposite effect of inhibiting COPI transport (Fig. 1b, Extended Data Fig. 1e). The expression levels of GAPDH in the different treatment conditions were also documented (Extended Data Fig. 1f). These initial results suggested that GAPDH acts as a negative regulator of COPI transport.

We next investigated COPI vesicle formation from Golgi membrane using a reconstitution system[6–8]. The addition of purified GAPDH to this in vitro assay inhibited vesicle formation (Fig. 1c). This effect was specific to GAPDH, as several other metabolic enzymes did not have a similar effect (Fig. 1c). We also performed electron microscopy, and found that GAPDH induced the accumulation of buds with constricted necks on the Golgi membrane (Fig. 1d). These results suggested that GAPDH inhibits COPI transport by targeting the fission stage of vesicle formation.

We next found that GAPDH binds directly to ARFGAP1 (Extended Data Fig. 1g) and interacts with ARFGAP1 in cells (Extended Data Fig. 1h). Thus, we performed a GAP assay, which showed that GAPDH—but not other metabolic enzymes—inhibits the catalytic activity of ARFGAP1 (Fig. 1e). Complementing this finding, we found that a mutant ARFGAP1 that is deficient in catalytic activity could not promote COPI vesicle fission (Fig. 1f). ARFGAP1 also acts as a coat component by promoting coat polymerization and cargo sorting[3,4]. However, GAPDH did not affect the interaction of ARFGAP1 with coatomer (Extended Data Fig. 1i) or with COPI cargo proteins (Extended Data Fig. 1j). Therefore, we concluded that GAPDH inhibits COPI vesicle fission by targeting the GAP activity of ARFGAP1. This role of GAPDH did not require its catalytic activity (Supplementary Information, and Extended Data Fig. 1k, l).

We then sought to understand how GAPDH could be recruited to the Golgi to inhibit COPI transport. Prompted by the observation that starvation redistributes GAPDH from the cytosol to the nucleus[9], we performed a subcellular fractionation experiment and found that starvation also redistributed GAPDH to cytoplasmic membranes (Extended Data Fig. 2a). Confocal microscopy revealed that this redistribution involved the Golgi (Extended Data Fig. 2b), the trans-Golgi network (TGN) (Extended Data Fig. 2c), early endosomes (Extended Data Fig. 2d) and the late endocytic compartments (Extended Data Fig. 2e). We also detected a pool of GAPDH at the endoplasmic reticulum (ER), but the level of this ER pool was not affected by starvation (Extended Data Fig. 2f).

We next examined whether GAPDH also inhibits other intracellular pathways. Performing a quantitative screen of the major pathways, as previously described[10], we found that siRNA against *GAPDH* did not affect transport from the ER to the Golgi (Fig. 2a, Extended Data Fig. 2g), but enhanced transport from the Golgi to the plasma membrane (Fig. 2b, Extended Data Fig. 2h). All three major endocytic routes were also enhanced, including endocytic recycling (Fig. 2c, Extended Data Fig. 2i), endocytic transport to the Golgi (Fig. 2d, Extended Data Fig. 2j) and endocytic transport to the lysosome (Fig. 2e, Extended Data Fig. 2k). Fluid-phase uptake was also enhanced by the siRNA treatment against *GAPDH* (Fig. 2f, Extended Data Fig. 2l). However, clathrin-mediated endocytosis (Fig. 2g) and two types of non-clathrin endocytosis (Extended Data Fig. 2m, n) were not affected. Moreover, similar to the case of COPI transport, we found that GAPDH overexpression had the opposite effect of inhibiting the affected pathways (Fig. 2a–g). The specificity of our detection of GAPDH and organelle markers were also confirmed (Supplementary Information and Extended Data Figs. 2o–r, 3a–l).

We then examined whether GAPDH inhibits these additional pathways also by targeting ARF GAPs. ARF6 and its cognate GAP, ACAP1, act in endocytic recycling[11]; GAPDH inhibited the GAP activity of ACAP1 towards ARF6 (Fig. 2h). ARF1 and its cognate GAP, AGAP1, act in endocytic transport to the lysosome[12]; GAPDH also inhibited the GAP activity of AGAP1 towards ARF1 (Fig. 2i). Moreover, other metabolic enzymes did not have similar effects (Extended Data Fig. 3m, n). In the case of exit from the ER, Sec23p is the cognate GAP for the ARF-like small GTPase Sar1p[13]. Consistent with our finding above that GAPDH does not inhibit this pathway, we found that GAPDH also did

---

[1]Division of Rheumatology, Immunology and Allergy, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA. [2]Division of Pulmonary and Critical Care Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA. [3]The FIRC Institute of Molecular Oncology, Milan, Italy. [4]Cardiovascular Division, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA. *e-mail: vhsu@bwh.harvard.edu

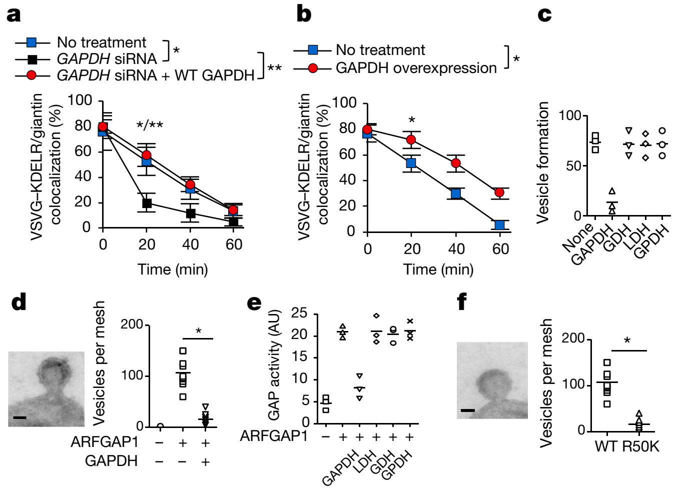


Fig. 1 | GAPDH inhibits COPI vesicle fission by targeting the GAP activity of ARFGAP1. a, **b**, COPI transport in HeLa cells. $^*P = 9.8 \times 10^{-7}$, $^{**}P = 9.2 \times 10^{-9}$ (**a**); $^*P = 6.8 \times 10^{-6}$ (**b**). **c**, Vesicle reconstitution system using Golgi membrane and purified proteins, with added metabolic enzymes as indicated. GDH, glutamate dehydrogenase; LDH, lactate dehydrogenase; GPDH, glycerol-3-phosphate dehydrogenase. **d**, Vesicle reconstitution system, with added purified proteins as indicated. Left, electron microscopy image of a Golgi membrane. Right, vesicle quantification. $^*P = 8.9 \times 10^{-7}$. **e**, GAP assay using ARF1 and ARFGAP1, with added metabolic enzymes as indicated. **f**, Vesicle reconstitution system, with added ARFGAP1 form as indicated. Left, electron microscopy image of Golgi membrane. Right, vesicle quantification. $^*P = 6.2 \times 10^{-6}$. In **a** and **b**, a representative experiment of three independent experiments is shown, data are mean ± s.d. of 10 fields of cells examined. In **c** and **e**, data are mean ± s.d. from a representative experiment of three independent experiments. In **d** and **f**, a representative experiment of three independent experiments is shown, data are mean ± s.d. of 10 electron microscopy meshes. *P* values determined by two-tailed *t*-test. Scale bars, 50 nm.

not inhibit the GAP activity of Sec23p towards Sar1p (Fig. 2j). GAPDH bound directly to ACAP1 (Extended Data Fig. 3o) and AGAP1 (Extended Data Fig. 3p), but not to Sec23p (Extended Data Fig. 3q). Therefore, these results suggest that GAPDH inhibits the transport pathways by targeting ARF GAPs. Moreover, led by the consideration that the catalytic domain in ARFGAP1 is conserved across many ARF GAPs, but not for Sec23p, we found that GAPDH bound to the catalytic domain of ARFGAP1 (Extended Data Fig. 3r), which suggests why GAPDH does not interact with Sec23p, and therefore does not inhibit exit from the ER. We also examined transferrin endocytosis and the results further supported that GAPDH inhibits the transport pathways by targeting ARF GAPs (Supplementary Information and Extended Data Figs. 3s–v).

We then examined the effect of starvation on the transport pathways. Cells were incubated with a general starvation medium (Fig. 3a–h), which lacks glucose and amino acids, Hank's medium (Extended Data Fig. 4a–h), which lacks amino acids and serum, or media that lack either glucose (Extended Data Fig. 4i–p) or amino acids (Extended Data Fig. 4q–x). In all cases, the same pathways were inhibited, and notably, these effects were reversed by siRNA against *GAPDH* (Fig. 3a–h and Extended Data Fig. 4a–x).

We then considered that vesicular transport consumes substantial cellular energy. We therefore hypothesized that the GAPDH-mediated transport inhibition enables the cell to conserve energy during starvation. In support of this hypothesis, we found that siRNA against *GAPDH* exacerbated the starvation-induced reduction in total ATP level, whereas GAPDH overexpression had the opposite effect (Fig. 3i, Extended Data Fig. 5a–c). Moreover, siRNA against *GAPDH* enhanced starvation-induced cell death, whereas GAPDH overexpression had the opposite effect (Fig. 3j). We ruled out that these effects could be attributed to the conventional role of GAPDH in glycolysis (Supplementary Information and Extended Data Fig. 5d–f). We also confirmed the key findings regarding how GAPDH acts in another cell type (Supplementary Information and Extended Data Fig. 5g–x).

We pursued another line of investigation that further supported GAPDH-mediated transport inhibition acting in energy homeostasis. AMPK has a central role in energy homeostasis, acting by sensing an energy deficit and then coordinating a wide range of cellular processes to restore energy homeostasis[14]. Treating cells with siRNA against *AMPKα1* (also known as *PRKAA1*), we found that starvation could no longer inhibit the transport pathways (Extended Data Fig. 6a–h); this was confirmed with a rescue experiment (Extended Data Fig. 6i, j), and with treatment using a different siRNA sequence (Extended Data Fig. 6k–s). Complementing these findings, we found that siRNA against

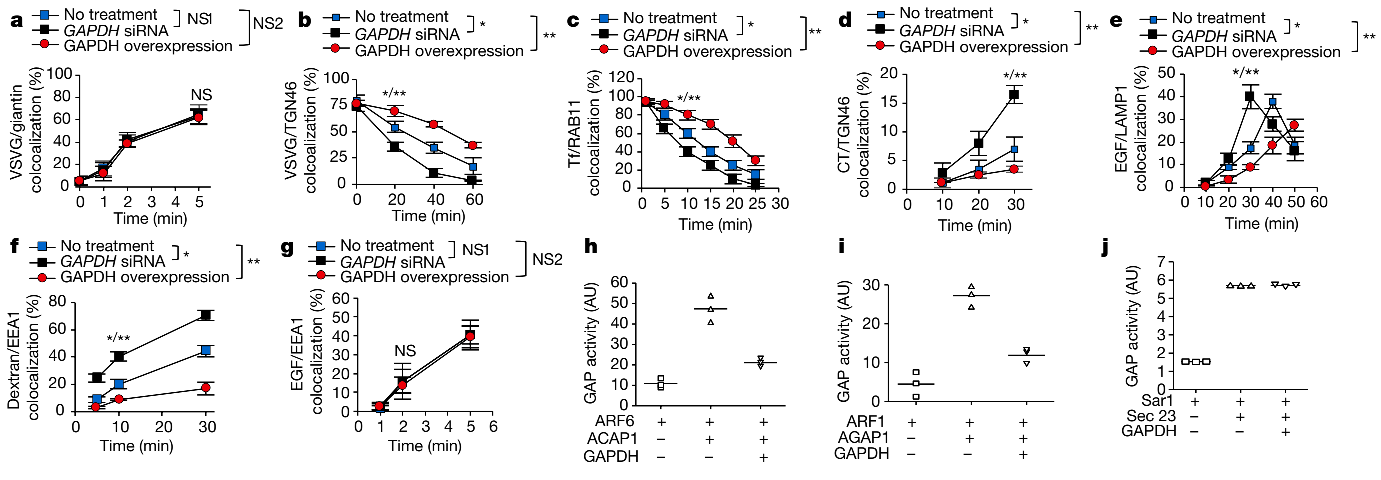


Fig. 2 | GAPDH inhibits other intracellular pathways also by targeting ARF GAPs. a–**g**, Transport assays in HeLa cells. **a**, Transport of vesicular stomatitis virus G protein (VSVG) from the ER to the Golgi. $P = 0.625$ (NS1); $P = 0.438$ (NS2). **b**, Transport of VSVG from the Golgi to the plasma membrane. $^*P = 7.6 \times 10^{-8}$, $^{**}P = 5.4 \times 10^{-6}$. **c**, Endocytic recycling of transferrin (Tf) from the early endosome to the plasma membrane. $^*P = 1.05 \times 10^{-7}$, $^{**}P = 1.05 \times 10^{-7}$. **d**, Endocytic transport of cholera toxin (CT) from the plasma membrane to the Golgi. $^*P = 2.7 \times 10^{-9}$, $^{**}P = 4.5 \times 10^{-4}$. **e**, Endocytic transport of epidermal growth factor (EGF) from the plasma membrane to the lysosome. $^*P = 1.6 \times 10^{-8}$, $^{**}P = 7.2 \times 10^{-7}$. **f**, Endocytosis of dextran. $^*P = 1.1 \times 10^{-10}$, $^{**}P = 1.8 \times 10^{-6}$. **g**, EGF endocytosis. $P = 0.9997$ (NS1); $P = 0.418$ (NS2). **h**–**j**, GAP assays using ARF6 and ACAP1 (**h**), ARF1 and AGAP1 (**i**), or Sar1p and Sec23p (**j**). In **a**–**g**, one representative experiment of three independent experiments is shown, data are mean ± s.d. of 10 fields of cells examined. In **h**–**j**, data are mean ± s.d. from a representative experiment of three independent experiments. *P* values determined by two-tailed *t*-test. NS, not significant.

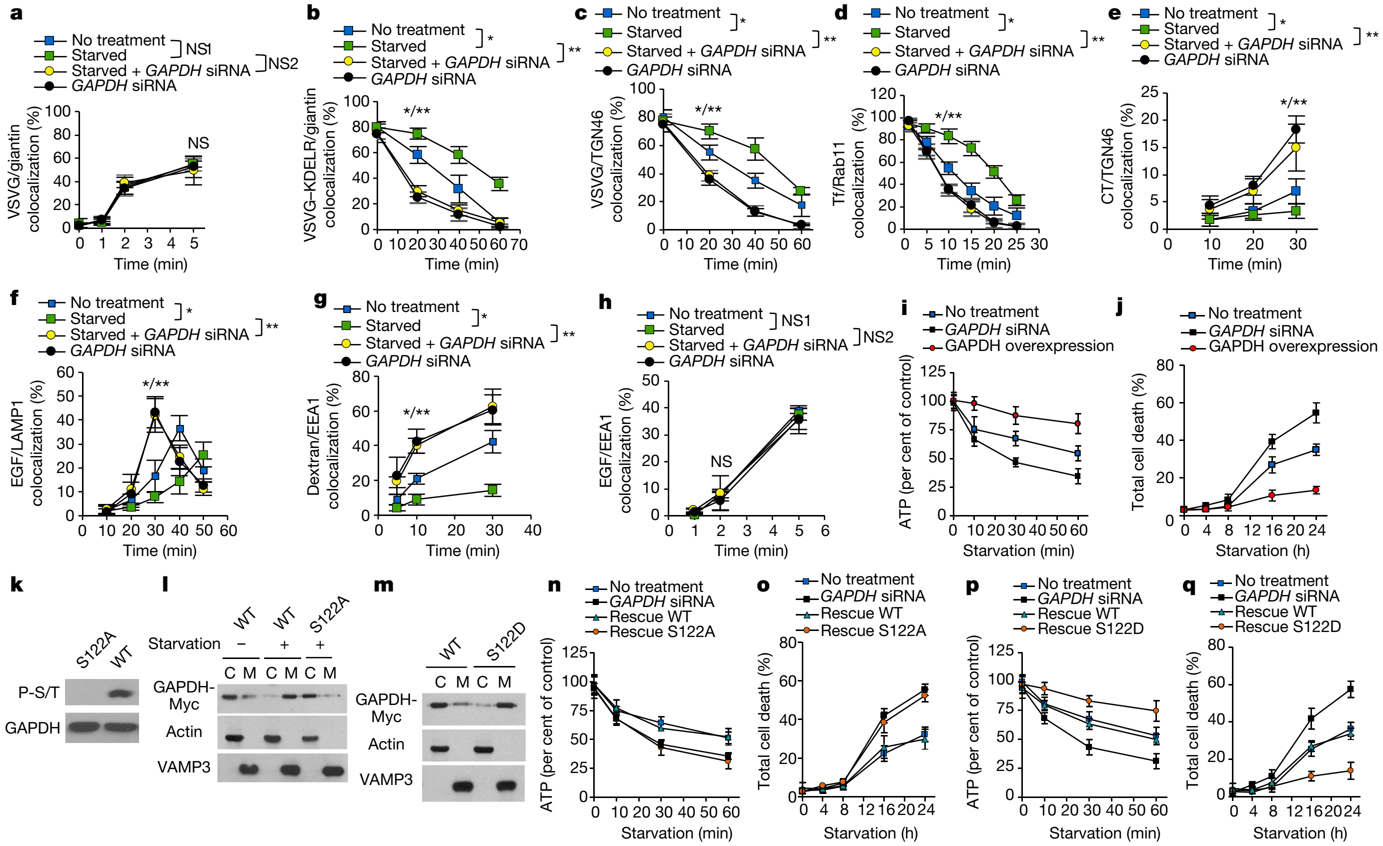


Fig. 3 | AMPK phosphorylates GAPDH to inhibit the transport pathways for energy homeostasis. Treatments as indicated. **a**–**h**, Transport assays in HeLa cells. **a**, ER to Golgi. $P = 0.7084$ (NS1), $P = 0.175$ (NS2). **b**, Golgi to ER. $^{*}P = 5.7 \times 10^{-6}$, $^{**}P = 6.3 \times 10^{-15}$. **c**, Golgi to plasma membrane. $^{*}P = 5.4 \times 10^{-6}$, $^{**}P = 4.3 \times 10^{-11}$. **d**, Early endosome to plasma membrane, $^{*}P = 5.5 \times 10^{-9}$, $^{**}P = 1.2 \times 10^{-12}$. **e**, Plasma membrane to Golgi. $^{*}P = 5.4 \times 10^{-4}$, $^{**}P = 6.2 \times 10^{-6}$. **f**, Plasma membrane to lysosome, $^{*}P = 1.7 \times 10^{-3}$, $^{**}P = 1.9 \times 10^{-8}$. **g**, Dextran endocytosis. $^{*}P = 7.8 \times 10^{-8}$, $^{**}P = 2.1 \times 10^{-12}$. **h**, EGF endocytosis. $P = 0.9994$ (NS1), $P = 0.7437$ (NS2). **i**, Total ATP level. **j**, Cell death. **k**, AMPK phosphorylation of GAPDH assessed by an in vitro kinase assay. Representative blot of three independent experiments. p-S/T, phosphoserine/threonine. **l**, **m**, Distribution of GAPDH in membranes (M) and cytosol (C). Representative gels of two independent experiments. **n**, **p**, Total ATP level. **o**, **q**, Cell death. In **a**–**h**, one representative experiment of three independent experiments is shown, data are mean ± s.d. of 10 fields of cells examined. In **i**, **j**, **n**, **p**, **o**, **q**, data are mean ± s.d. from a representative experiment of three independent experiments. *P* values determined by two-tailed *t*-test.

AMPKα1 prevented starvation from redistributing GAPDH to membrane compartments (Extended Data Figs. 7a–e).

We also activated AMPK in the normal (non-starved) condition by treating cells with 5-aminoimidazole-4-carboxamide ribonucleotide (AICAR), and found that it reproduced the effect of starvation in suppressing the identical transport pathways, and this effect was also reversed by siRNA against *GAPDH* (Extended Data Fig. 7f–m). This result was further confirmed using A769662, a more specific activator of AMPK (Extended Data Fig. 7n–u). Moreover, AICAR treatment in the normal condition was sufficient to redistribute GAPDH to membrane compartments (Extended Data Fig. 8a–e). Thus, the collective results suggested that starvation acts through AMPK and then GAPDH to inhibit the transport pathways.

To examine how AMPK affects GAPDH, we were initially led by a previous observation that AMPK phosphorylates GAPDH at serine 122[9]. We confirmed this finding, observing that AMPK did not phosphorylate GAPDH(S122A) (Fig. 3k, Extended Data Fig. 8f, g). GAPDH was phosphorylated less efficiently when compared to acetyl-CoA carboxylase, a well-established substrate of AMPK (Extended Data Fig. 8h–k), possibly because the sequence surrounding the S122 residue does not represent an optimal consensus site for AMPK phosphorylation[15]. We also pursued cell-based studies, confirming that starvation activates AMPK in cells (Extended Data Fig. 8l). Starvation also increased the phosphorylation of GAPDH S122, and this effect was prevented by the S122A mutation (Extended Data Fig. 8m). Moreover, this phosphorylation was prevented by siRNA against *AMPKα1* (Extended Data Fig. 8n). Similar results were also seen in HEK293 cells (Extended Data Fig. 8o–q). Starvation also did not redistribute cytosolic GAPDH(S122A) to membranes (Fig. 3l). Consistent with this result, GAPDH(S122A) also prevented starvation from inhibiting the transport pathways (Extended Data Fig. 9a–h). When GAPDH S122 was mutated to aspartate (S122D), which mimics constitutive phosphorylation at this site[9], we found that simply expressing GAPDH(S122D) in the non-starved condition was sufficient to redistribute cytosolic GAPDH to membranes (Fig. 3m). The expression of this mutant in the normal condition was also sufficient to inhibit the transport pathways (Extended Data Fig. 9i–p).

The effects of mutating the S122 residue in GAPDH suggested yet another way of confirming that GAPDH-mediated inhibition of transport acts in energy homeostasis. Replacing the endogenous GAPDH with the S122A mutant in cells exacerbated the decline in both the total ATP level and cell viability induced by starvation (Fig. 3n, o). By contrast, replacing endogenous GAPDH with the S122D mutant had the opposite effect of increasing the ATP level (Fig. 3p) and cell viability (Fig. 3q) during starvation.

AMPK phosphorylation of the S122 residue in GAPDH has previously been found to induce the translocation of GAPDH to the nucleus, where it stimulates sirtuin 1 (SIRT1) activity to activate a transcription program that promotes autophagy[9]. We confirmed that our general starvation medium (which lacks glucose and amino acids) activated autophagy (Fig. 4a, b), and found that the recruitment of GAPDH to the Golgi—which is needed to inhibit COPI transport—began immediately

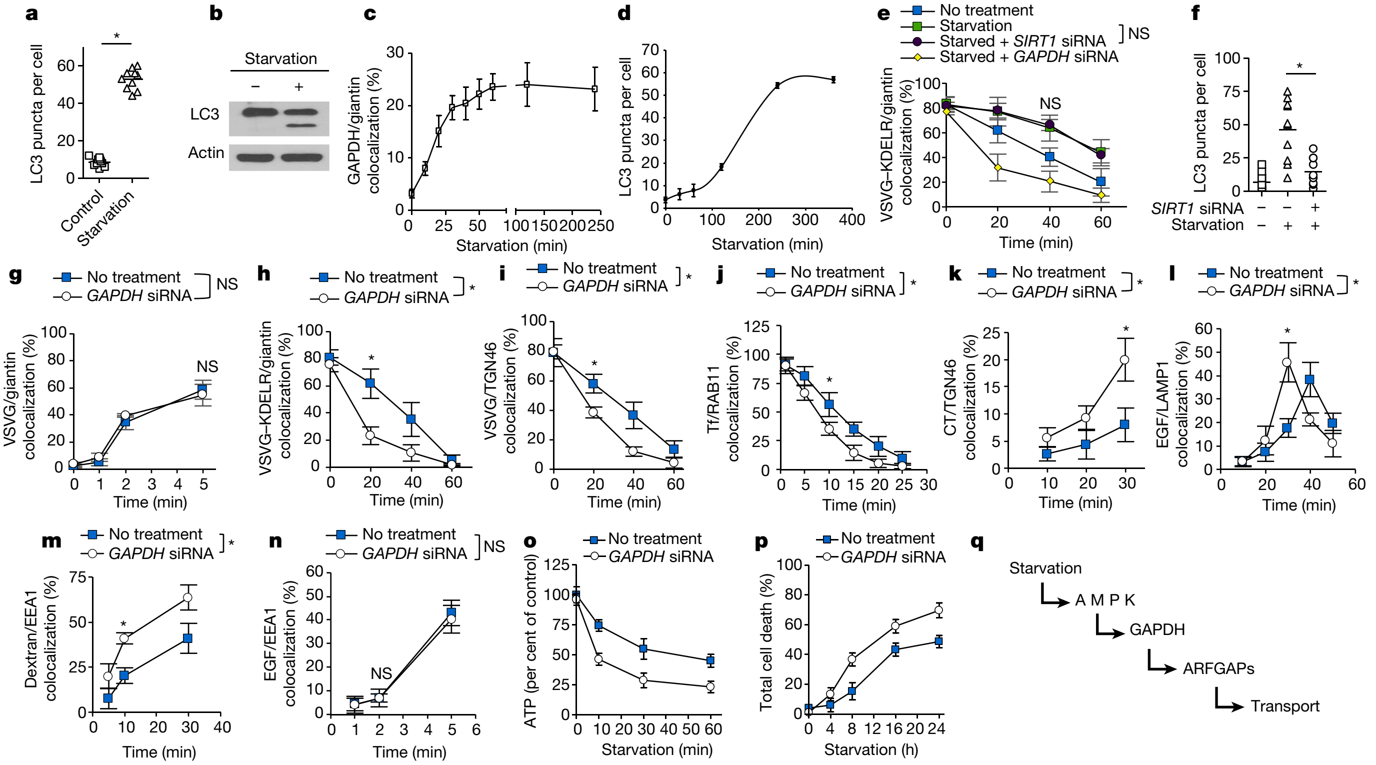


Fig. 4 | The roles of GAPDH in transport and autophagy are distinct. **a**, **b**, Autophagy in HeLa cells following starvation, assessed by LC3 puncta formation (**a**; *$P = 0.000152$) or LC3 gel shift (**b**). **c**, GAPDH recruitment to the Golgi in HeLa cells with the indicated conditions. **d**, LC3 puncta formation in HeLa cells. **e**, COPI transport assay in HeLa cells. $P = 0.8490$ (NS). **f**, LC3 puncta formation in HeLa cells. *$P = 0.00174$. **g**–**n**, Transport assays in ATG5-deficient MEFs. **g**, ER to Golgi. $P = 0.25$ (NS). **h**, Golgi to ER. *$P = 9.9 \times 10^{-8}$. **i**, Golgi to plasma membrane. *$P = 5.9 \times 10^{-7}$. **j**, Early endosome to plasma membrane. *$P = 6.4 \times 10^{-5}$. **k**, Plasma membrane to Golgi. *$P = 7.6 \times 10^{-7}$. **l**, Plasma membrane to lysosome. *$P = 3.2 \times 10^{-7}$. **m**, Dextran endocytosis. *$P = 1.2 \times 10^{-9}$. **n**, EGF endocytosis. $P = 0.9994$ (NS). **o**, Total ATP level. **p**, Cell death. **q**, Summary of how inhibition of transport by GAPDH promotes cellular energy homeostasis. A representative experiment of one (**d**), two (**a**, **f**) or three (**e**, **g**–**n**) independent experiments is shown, data are mean ± s.d. of 10 fields of cells examined. In **b**, the gel is representative of two independent experiments. In **c**, data are mean ± s.d. of 6 fields of cells examined in one experiment. In **o** and **p**, data are mean ± s.d. from three independent experiments. *P* values determined by two-tailed *t*-test.

upon starvation (Fig. 4c). By contrast, LC3 puncta formation, which tracks autophagosome formation, started more slowly (Fig. 4d). Further distinguishing between the roles of GAPDH in transport inhibition and autophagy, we found that siRNA against *SIRT1* did not have an appreciable effect on starvation-induced inhibition of COPI transport (Fig. 4e), but prevented starvation from inducing autophagy (Fig. 4f).

We also examined mouse embryonic fibroblasts (MEFs) derived from Atg5-deficient mice, which cannot undergo autophagy[16], and found that GAPDH could still inhibit the identical transport pathways as those seen in autophagy-competent cells (Fig. 4g–n). Moreover, siRNA against *GAPDH* could still exacerbate the decline of total ATP (Fig. 4o) and cell viability (Fig. 4p) during starvation. In another approach, we modified GAPDH so that its redistribution to the nucleus was prevented, but its recruitment to cytoplasmic membrane compartments was preserved. This form of GAPDH could still inhibit the transport pathways, and increased ATP level and cell viability during starvation (Supplementary Information and Extended Data Fig. 10a–u). Therefore, multiple lines of evidence all point to the roles of GAPDH in transport inhibition and autophagy being distinct.

In summary, we have defined a new mechanism of energy homeostasis that is critical for the cell to survive during starvation. Mechanistically, this involves the activation of AMPK to induce the redistribution of cytosolic GAPDH to multiple membrane compartments, which then targets ARF GAPs to inhibit different transport pathways (Fig. 4q). As the cell encounters starvation in many settings, considerable effort has been devoted to elucidating mechanisms of cell survival in this condition. Autophagy has emerged as a major mechanism[1]. In the overall energy equation, autophagy acts to increase energy availability. By contrast, transport inhibition by GAPDH acts to reduce energy consumption. These fundamentally distinct mechanisms complement each other to promote energy homeostasis during starvation. Another notable distinction between the two mechanisms is that the transport inhibition by GAPDH occurs more rapidly than autophagy. This more rapid mechanism is likely to enable the cell to stave off the lethal consequences of starvation while longer-term solutions, such as autophagy, are being mobilized.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0475-6.

Received: 17 November 2016; Accepted: 12 July 2018;
Published online 12 September 2018.

1. Singh, R. & Cuervo, A. M. Autophagy in the cellular energetic balance. *Cell Metab.* **13**, 495–504 (2011).
2. Donaldson, J. G. & Jackson, C. L. ARF family G proteins and their regulators: roles in membrane transport, development and disease. *Nat. Rev. Mol. Cell Biol.* **12**, 362–375 (2011).
3. Yang, J. S. et al. ARFGAP1 promotes the formation of COPI vesicles, suggesting function as a component of the coat. *J. Cell Biol.* **159**, 69–78 (2002).
4. Lee, S. Y., Yang, J. S., Hong, W., Premont, R. T. & Hsu, V. W. ARFGAP1 plays a central role in coupling COPI cargo sorting with vesicle formation. *J. Cell Biol.* **168**, 281–290 (2005).
5. Sirover, M. A. On the functional diversity of glyceraldehyde-3-phosphate dehydrogenase: biochemical mechanisms and regulatory control. *Biochim. Biophys. Acta* **1810**, 741–751 (2011).
6. Yang, J. S. et al. A role for phosphatidic acid in COPI vesicle fission yields insights into Golgi maintenance. *Nat. Cell Biol.* **10**, 1146–1153 (2008).
7. Yang, J. S. et al. COPI acts in both vesicular and tubular transport. *Nat. Cell Biol.* **13**, 996–1003 (2011).

8. Park, S. Y., Yang, J. S., Schmider, A. B., Soberman, R. J. & Hsu, V. W. Coordinated regulation of bidirectional COPI transport at the Golgi by CDC42. *Nature* **521**, 529–532 (2015).
9. Chang, C. et al. AMPK-dependent phosphorylation of GAPDH triggers Sirt1 activation and is necessary for autophagy upon glucose starvation. *Mol. Cell* **60**, 930–940 (2015).
10. Bai, M. et al. ARFGAP1 promotes AP-2 dependent endocytosis. *Nat. Cell Biol*. **13**, 559–567 (2011).
11. Li, J. et al. An ACAP1-containing clathrin coat complex for endocytic recycling. *J. Cell Biol*. **178**, 453–464 (2007).
12. Nie, Z. et al. Specific regulation of the adaptor protein complex AP-3 by the ARF GAP AGAP1. *Dev. Cell* **5**, 513–521 (2003).
13. Barlowe, C. et al. COPII: a membrane coat formed by Sec proteins that drive vesicle budding from the endoplasmic reticulum. *Cell* **77**, 895–907 (1994).
14. Hardie, D. G., Ross, F. A. & Hawley, S. A. AMPK: a nutrient and energy sensor that maintains energy homeostasis. *Nat. Rev. Mol. Cell Biol*. **13**, 251–262 (2012).
15. Schaffer, B. E. et al. Identification of AMPK phosphorylation sites reveals a network of proteins involved in cell invasion and facilitates large-scale substrate prediction. *Cell Metab*. **22**, 907–921 (2015).
16. Kuma, A. et al. The role of autophagy during the early neonatal starvation period. *Nature* **432**, 1032–1036 (2004).

Acknowledgements This work was supported by grants from the US National Institutes of Health to V.W.H. (GM058615 and GM115683), J.Lo. (HL61795, HL119145, HG007690 and GM107618) and W.M.O. (HL128802), and a grant from the American Heart Association to J.Lo. (D700382).

Reviewer information *Nature* thanks D. Hardie and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions J.-S.Y., J.-W.H., S.-Y.P. and J.Li performed colocalization studies. J.-S.Y. performed vesicle reconstitution studies with help from G.V.B. and A.A.M. J.-W.H. performed membrane fractionation studies. J.-S.Y. and J.-W.H. performed ATP and cell death assays. W.M.O. performed metabolic studies with supervision by J.Lo. All authors participated in experimental design and data analysis. V.W.H. supervised the project and wrote the manuscript with input from all other authors.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0475-6.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0475-6.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to V.W.H.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Chemicals, proteins and cells. AICAR, chloroquine diphosphate, methyl pyruvate, protease inhibitor cocktail and phosphatase inhibitor cocktail were obtained from Sigma. AMPK activator A-769662 was obtained from Cayman. Alexa Fluor 546-conjugated transferrin, and Alexa Fluor 555-conjugated forms of EGF, dextran and cholera toxin were obtained from Invitrogen.

Purification of coatomer, ARF1, ARFGAP1, BARS, Golgi-enriched membrane and cytosol have been described[3,17]. Purification of ARF6, ACAP1, and AGAP1 have also been described[11,12]. Purified mammalian Sar1p and Sec23p were obtained from J. Goldberg, Memorial Sloan-Kettering Cancer Center. Purified forms of GAPDH, glutamate dehydrogenase (GDH), glycerol-3-phosphate dehydrogenase (GPDH), and lactate dehydrogenase (LDH) were obtained from Sigma. AMPK and the SAMS peptide (HMRSAMSGLHLVKRR) were obtained from (Promega).

For the purification of recombinant GAPDH, the pET30-2-GAPDH plasmid was transformed into BL21 competent cells. Isopropyl β-D-1-thiogalactopyranoside (0.1mM) was added to induce protein expression. Bacteria were then resuspended in 5 ml lysis buffer (20 mM Tris pH 8.0, 100 mM NaCl, 10 mM imidazole, 1 mg/ml lysozyme, and protease inhibitors cocktail). After gentle agitation at 4 °C for 30 min, the suspension was sonicated on ice 6 times for 10 s each, and then centrifuged at 12,000*g* for 20 min. The supernatant was recovered and then added to 0.5 ml Ni-beads (50% slurry), followed by incubation at 4 °C for 1 h. Beads were then washed with 10 ml washing buffer (20 mM Tris pH 8.0, 100 mM NaCl, 10 mM imidazole). Bound proteins were then eluted with elution buffer (20 mM Tris pH 8.0, 100 mM NaCl, 250 mM imidazole).

HeLa, BSC-1, COS-7, CHO and HEK293 cells have previously been described[10,18,19]. Wild-type and ATG5-deficient MEF cells were obtained from N. Mizushima, University of Tokyo. Cells were cultured in Dulbecco's Modified Eagle Medium (DMEM) with 10% fetal bovine serum and supplemented with glutamine. Cells had been free of mycoplasma contamination upon their receipt, but were not tested immediately prior to use. Unless stated otherwise, starvation involves incubating cells in a medium consisting of phosphate-buffered saline (PBS) with 10% dialysed bovine serum.

Mass spectrometry. Cytosol was purified from rat liver by homogenization in lysis buffer (1 mM Tris pH 7.4, 800 mM sucrose, 5 mM EDTA and protease inhibitors) at 4 °C, followed by centrifugation at 80,000*g* for 90 min. The resulting supernatant was dialysed against dialysis buffer (25mM Tris at pH 8.0, 50mM KCl and 1mM DTT), followed by centrifugation at 150,000*g* for 90 min at 4 °C. Aliquots were stored at −80 °C. Pull-down experiments were performed by incubating glutathione *S*-transferase (GST) or GST–ARFGAP1 on beads (5 μg) with 50 μg rat liver cytosol in traffic buffer (25 mM HEPES pH 7.2, 50 mM KCl and 2.5 mM Mg(OAc)_2) for 1 h followed by extensive washing. Samples were separated by SDS–PAGE and then stained with Coomassie blue, followed by gel excision for protein identification by mass spectrometry (Taplin Mass Spectrometry Facility, Harvard Medical School). Those proteins (having at least five identified peptides) that interacted with GST–ARFGAP1 but not GST were considered specific.

Plasmids and antibodies. Mammalian expression vectors that contain temperature-sensitive mutants (ts-045) of VSVG and VSVG–KDELR have been described[7]. GAPDH for mammalian expression was subcloned into pcDNA3.1 and pEGFP-N1. GAPDH for bacterial expression (pET30-2-GAPDH) was obtained from Addgene (83910). Mutant forms of GAPDH were generated using QuikChange Site-Directed-Mutagenesis (Stratagene). The sequences of oligonucleotides used for mutagenesis are listed in Supplementary Table 1. The pEGFP-NES-GAPDH construct was obtained from W. Liu (Zhejiang University). IL2R-β tagged with haemagglutinin (HA) and Venus (HA-Venus-IL2R-β) was obtained from T. Waldman and O. Hurtaldo (NIH). GST-ARFGAP1, GST-Wbp1 and GST-KDELR have been described[3,17].

Antibodies against ARF1, ARFGAP1, β-COP, calnexin, cellubrevin, coatomer, giantin, GM130, LAMP1, Sec61p, TGN46 VSVG, TGN46, HA tag and Myc tag have previously been described[10,17,19]. The following antibodies were also used: AMPKα (Cell Signaling, 2603S), β-actin (Ambion, AM4302), GFP (Invitrogen, GF28R), LC3 (Sigma, L7543), p62 (Santa Cruz, 28359), p172-AMPKα (Cell Signaling, 2531S), RAB11 (BD Biosciences, 610656), SIRT1 (Cell Signaling, 2493S), and Flag tag M2 (Sigma, F1804). Antibodies against GAPDH were obtained from three sources: Ambion (AM4300), Santa Cruz (FL335), and Sigma (G9545). Anti-pS/pT was obtained from Abcam (117253) and conjugated antibodies were obtained from Jackson ImmunoResearch, including Cy2- or Cy3-conjugated donkey antibodies against mouse or rabbit IgG, and horseradish peroxidase-conjugated donkey antibodies against mouse or rabbit IgG.

Transfections and siRNA. Transfection of DNA plasmids was performed using FuGene6 (Roche). Transfection of siRNA was performed using Lipofectamine RNAiMAX (Invitrogen). Sequences used for siRNA against *GAPDH* and *AMPKα1*, and primers used to generate siRNA-resistant forms of *GAPDH* and *AMPKα1* are listed in Supplementary Table 1. Rescue plasmids were generated using QuikChange Site-Directed Mutagenesis (Stratagene).

We found that siRNA against *GAPDH* for 2 days was optimal for reducing the endogenous level while maintaining cell viability. Rescue of siRNA against *GAPDH* involved the transfection of siRNA-resistant *GAPDH* for 12 h to achieve limited GAPDH expression. GAPDH overexpression involved the transfection of GAPDH for 48 h.

In vivo transport assays. A quantitative microscopy-based approach, which involves the colocalization of model cargoes with organelle markers coupled with kinetic analysis was performed as previously described[8,10].

For anterograde transport from ER to Golgi, cells were transfected with pROSE-VSVG-ts045-Myc for 1 day, and then incubated at 39 °C for 4 h to accumulate VSVG in the ER. Cells were then shifted to 32 °C for different times as indicated in figures. Cells were then stained for giantin, followed by confocal microscopy to assess the arrival of VSVG to the Golgi.

For retrograde transport from the Golgi to the ER, cells were transfected with pROSE-VSVG-ts045-KDELR-Myc for 1 day, and then incubated at 32 °C for 8 h to achieve steady-state distribution at the Golgi. Cells were then shifted to 39 °C for different times as indicated in the figures. Cells were then stained for giantin, followed by confocal microscopy to assess the exit of VSVG-KDELR from the Golgi.

For anterograde transport from the Golgi to the plasma membrane, cells were transfected with pROSE-VSVG-ts045-Myc for 1 day, and then incubated at 20 °C for 2 h to accumulate VSVG at the TGN. Cells were then shifted to 32 °C for different times as indicated in the figures. Cells were then stained for TGN46, followed by confocal microscopy to assess the exit of VSVG from the Golgi.

For the recycling of transferrin from the early recycling endosome to the plasma membrane, cells were incubated with Alexa 546-conjugated transferrin (5 μg/ml in DMEM) at 37 °C for 2 h to allow the steady-state accumulation of transferrin in endosomes. Subsequently, cells were incubated with medium without transferrin for different times as indicated in the figures. Cells were then stained for RAB11, followed by confocal microscopy to assess the exit of transferrin from the early recycling endosome.

For the retrograde transport of cholera toxin from the plasma membrane to the Golgi, cells were incubated with Alexa 555-conjugated cholera toxin for 30 min at 4 °C (0.5 μg/ml in DMEM). After washing to clear unbound cholera toxin, cells were shifted to 37 °C for different times as indicated in the figures. Cells were then stained for TGN46, followed by confocal microscopy to assess the arrival of cholera toxin to the Golgi.

For endocytic transport of EGF to the lysosome, cells were incubated with Alexa 555-conjugated EGF (1 μg/ml in DMEM) for 1 h at 4 °C. Cells were then washed to clear unbound EGF, followed by shifting to 37 °C for times indicated in the figures. Cells were stained for LAMP1, followed by confocal microscopy to assess the arrival of EGF to the lysosome.

For the fluid-phase uptake of dextran, cells were incubated with Alexa 555-conjugated dextran (0.2 mg/ml) at 37 °C for different times as indicated in the figures. Cells were then stained for EEA1, followed by confocal microscopy to assess the arrival of dextran to the early endosome.

For the endocytosis of EGF, cells were incubated with Alexa 555-conjugated EGF (1 μg/ml in DMEM) for 1 h at 4 °C. Cells were then washed to clear unbound EGF, followed by shifting to 37 °C for times indicated in the figures. Cells were stained for EEA1, followed by confocal microscopy to assess the arrival of EGF to the early endosome.

For the endocytosis of transferrin, cells were incubated with Alexa 546-conjugated transferrin (5 μg/ml in DMEM) for 1 h at 4 °C. Cells were then washed to clear unbound transferrin, followed by shifting to 37 °C for times indicated in the figures. Cells were stained for EEA1, followed by confocal microscopy to assess the arrival of transferrin to the early endosome.

For the endocytosis of cholera toxin, cells were incubated with Alexa 555-conjugated cholera toxin for 30 min at 4 °C (0.5 μg/ml in DMEM). Cells were then washed to clear unbound cholera toxin, followed by shifting to 37 °C for different times as indicated in the figures. Cells were stained for EEA1, followed by confocal microscopy to assess the arrival of cholera toxin to the early endosome.

For the endocytosis of IL2R-β, cells were transfected with pCDH-HA-Venus-IL2R-β for 1 day at 37 °C. Cells were then incubated with anti-HA antibody at 4 °C for 1 h to label the surface pool of the receptor. Cells were washed to clear unbound antibody, followed by shifting to 37 °C for different times as indicated in the figures. Cells were then stained for EEA1, followed by confocal microscopy to assess the arrival of IL2R-β to the early endosome.

Confocal microscopy. Colocalization studies were performed with either a Nikon or a Zeiss system. The Nikon system is equipped with the Nikon Eclipse TE2000U Inverted Microscope having a Plan Apo 60×/1.40 oil objective, Nikon D-Eclipse C1 confocal package with a 488 Laser (having 515/30 emission filter) and a 543 Laser (having 590/50 emission filter), and Nikon EZ-C1 v.3.90 acquisition software. The Zeiss system is equipped with the Zeiss Axio Observer Z1 Inverted Microscope having a Plan-Apochromat 63× objective, the Zeiss LSM 800 with Airyscan confocal package with Zeiss URGB (488- and 561-nm) laser lines, and Zen 2.3 blue edition confocal acquisition software.

For quantification of colocalization, ten fields of cells were examined, with each field typically containing about 5 cells. Images were imported into the NIH ImageJ v.1.50e software, and then analysed through a plugin software (https://imagej.net/Coloc_2). Under the 'image' tab, the 'split channels' option was selected. Under the 'plugins' tab, 'colocalization analysis' option was selected, and within this option, the 'colocalization threshold' option was selected. Colocalization values were then calculated by the software, and expressed as the fraction of protein of interest (cargo or GAPDH) colocalized with an organelle marker.

In vitro reconstitution of COPI vesicle formation. The reconstitution was performed essentially as described[17]. In brief, Golgi membrane (0.2 mg/ml) was washed with 3 M KCl, and then incubated with ARF1 (6 μg/ml) and coatomer (6 μg/ml) for the first-stage incubation that reproduces the ARF-dependent recruitment of coatomer onto Golgi membrane. The Golgi membrane was re-isolated and then incubated with ARFGAP1 (2 μg/ml), BARS (2 μg/ml) for the second stage that results in vesicle formation. GAPDH (1 μg/ml) was added at the second stage. Electron microscopy examination of Golgi membrane using the whole-mount technique has previously been described[3]. Quantification involves the examination of 10 meshes per condition.

Fractionation of the cytoplasm to obtain total membrane versus cytosol. Cells were washed with PBS, resuspended in homogenization buffer (0.25 M sucrose, 1 mM EDTA, and 20 mM HEPES-KOH, pH 7.4 and protease inhibitor cocktail) and then disrupted by passing through 28-gauge needles. After low-speed centrifugation (800*g* for 6 min) to spin out nuclei and unbroken cells, the resulting post-nuclear supernatant was centrifuged at 100,000*g* for 1 h to obtain cytosol and total membrane fractions.

Other in vitro assays. For pull-down assays, GST fusion proteins were bound to glutathione beads, then incubated with purified proteins. For the GAP assay, recombinant forms of ARF small GTPase were first loaded with GTP and then incubated with different ARF GAPs in the presence of liposomes. GTP hydrolysis was then quantified by detecting the generation of free phosphate (ATPase/GTPase Activity Assay Kit, Sigma). The in vitro kinase assay was performed by incubating AMPK (0.1 μg) with the SAMS peptide (a sequence derived from acetyl-CoA carboxylase) (3 μg) or different GAPDH proteins (40 μg) in 50 μl reaction buffer (25 mM MOPS, pH 7.2, 12.5 mM β-glycero-phosphate, 25 mM $MgCl_2$, 5 mM EGTA, 2 mM EDTA, 0.25 mM DTT, 100 μM AMP and 50 μM ATP) at 30 °C for times as indicated. Phosphorylation of the substrate was then quantified using the ADP-Glo Kinase assay (Promega), which detects ADP generated from ATP in the kinase reaction.

Other in vivo assays. Cellular ATP level was detected using ATPlite Luminescence Detection Assay System (Perkin Elmer), and was performed according to the manufacturer's instructions, with final values normalized for cell number. Cell death was assessed by incubating cells with propidium iodide (1 ng/μl), and then quantifying for positively stained cells by flow cytometry. Lactate level was measured using a colorimetric assay kit (Biovision). Oxygen consumption rate (OCR) was measured using an XFe24 Extracellular Flux Analyzer (Seahorse Biosciences), as previously described[20]. SIRT1 activity was measured using SIRT1 Direct Fluorescent Screening Assay Kit (Cayman), as previously described[9]. Autophagy was assessed through LC3 lipidation, using either confocal microscopy to quantify LC3 puncta formation, or western blotting to detect LC3 mobility shift, as previously described[9].

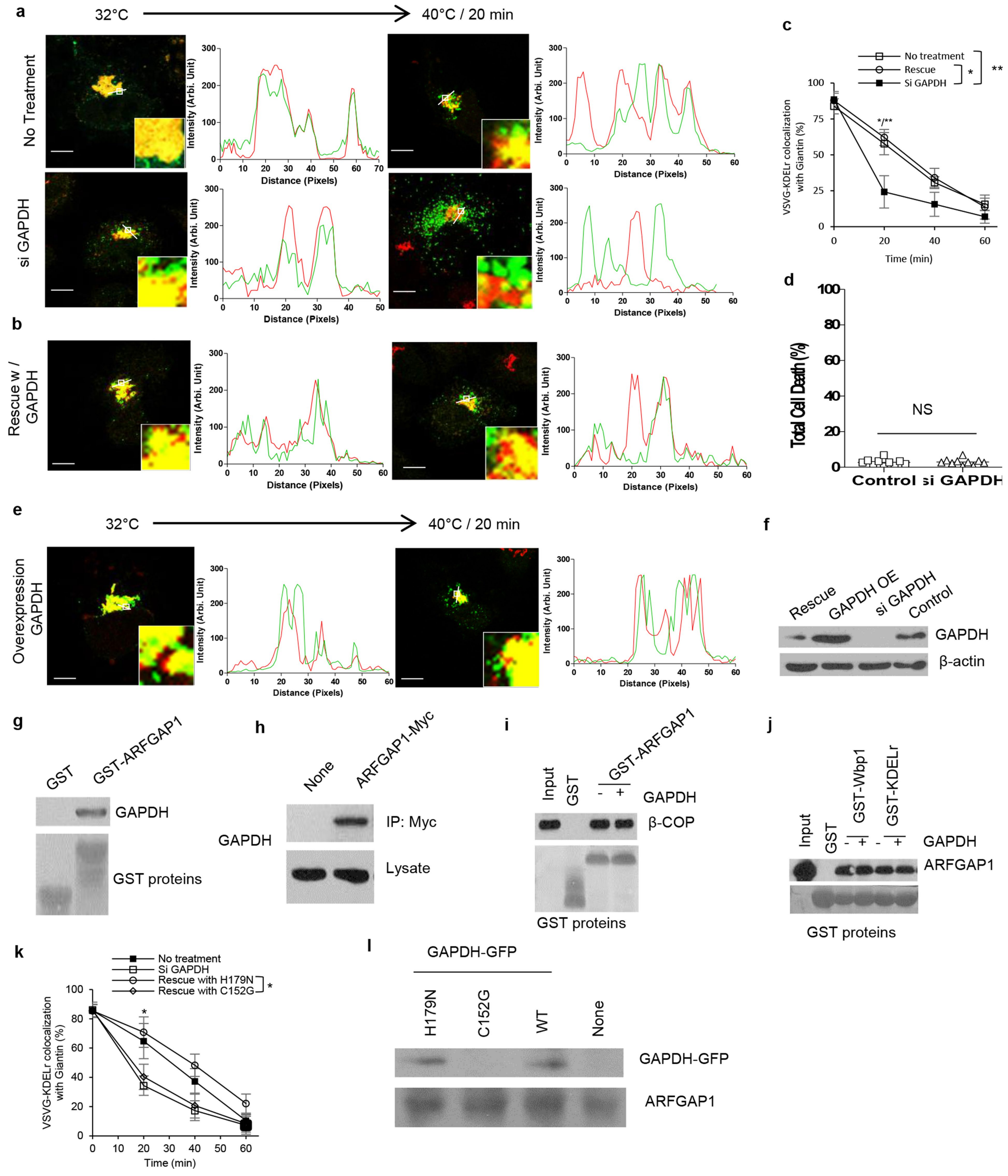
Experimental replicates. All colocalization studies are shown as 10 technical replicates from a representative experiment, with 3 independent experiments having been performed. Metabolic studies involving oxygen consumption rate are performed as 5 technical replicates, with 2 independent experiments having been performed. The result from a representative replicate is shown. All other studies show a representative experiment with the number of independent experiments indicated in the figure legend.

Statistical analysis. Sample size is noted in the figure legends. No statistical methods were used to predetermine sample size. Statistical significance was determined using Excel or Prism software for the two-tailed Student's *t*-test. No inclusion or exclusion criteria were pre-established. The experiments were not randomized. The investigators were not blinded to the group allocation during experiments and in outcome assessment.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

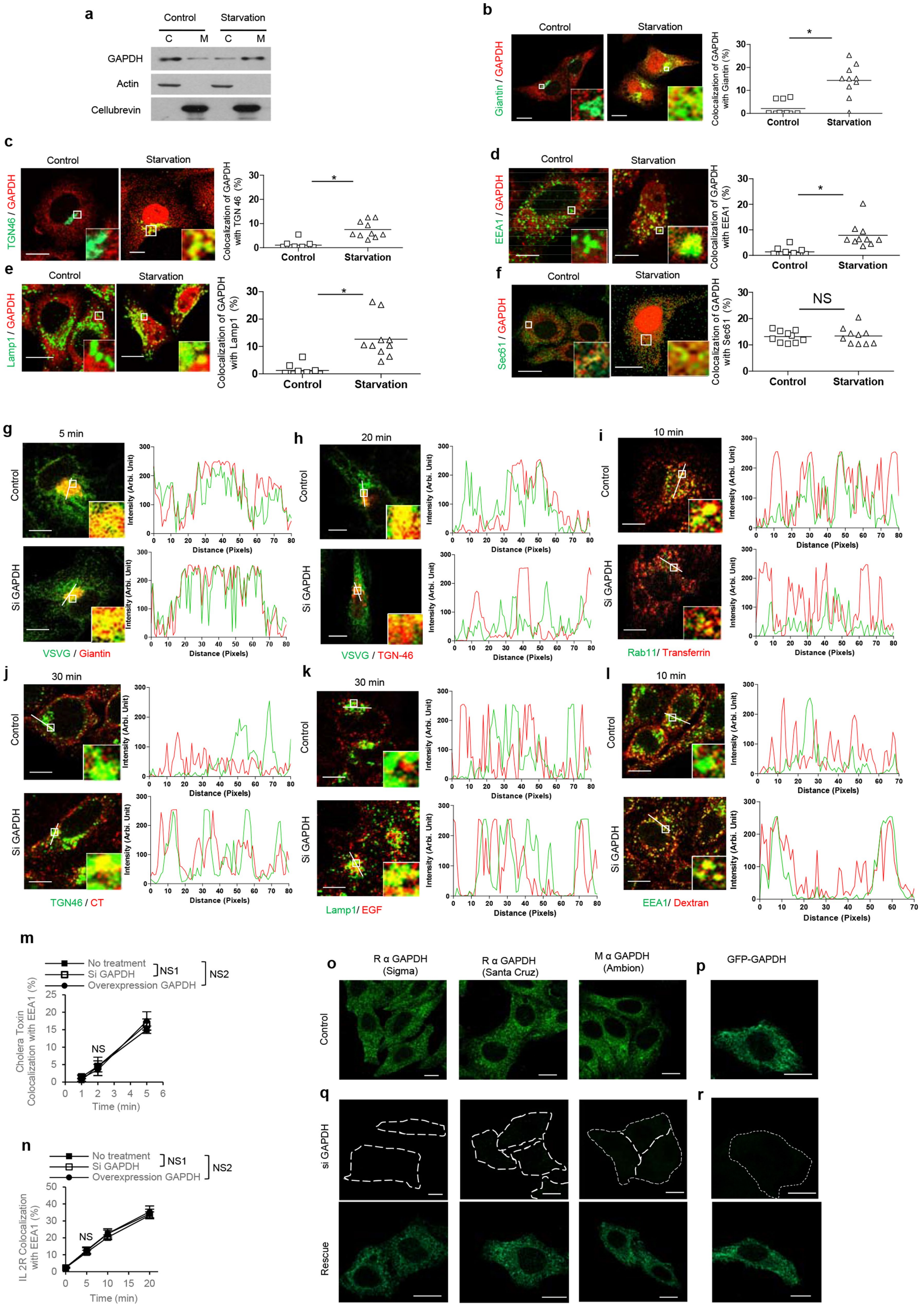
Data availability.. The data that support the findings of this study are available from the corresponding author upon reasonable request. The following figures have associated raw data: Figs. 1a, b, 2a–g, 3a–j, n–q, 4c–e, g–p; Extended Data Figs. 1c, k, 2m,n, s–v, 3 s–v, 4a–x, 5a–c, g–x, 6a–h, j, l–s, 7f–u, 9a–p, 10h–u. For gel source data, see Supplementary Fig. 1.

17. Yang, J. S. et al. A role for BARS at the fission step of COPI vesicle formation from Golgi membrane. *EMBO J.* **24**, 4133–4143 (2005).
18. Aoe, T. et al. The KDEL receptor, ERD2, regulates intracellular traffic by recruiting a GTPase-activating protein for ARF1. *EMBO J.* **16**, 7305–7316 (1997).
19. Dai, J. et al. ACAP1 promotes endocytic recycling by recognizing recycling sorting signals. *Dev. Cell* **7**, 771–776 (2004).
20. Oldham, W. M., Clish, C. B., Yang, Y. & Loscalzo, J. Hypoxia-mediated increases in l-2-hydroxyglutarate coordinate the metabolic response to reductive stress. *Cell Metab.* **22**, 291–303 (2015).



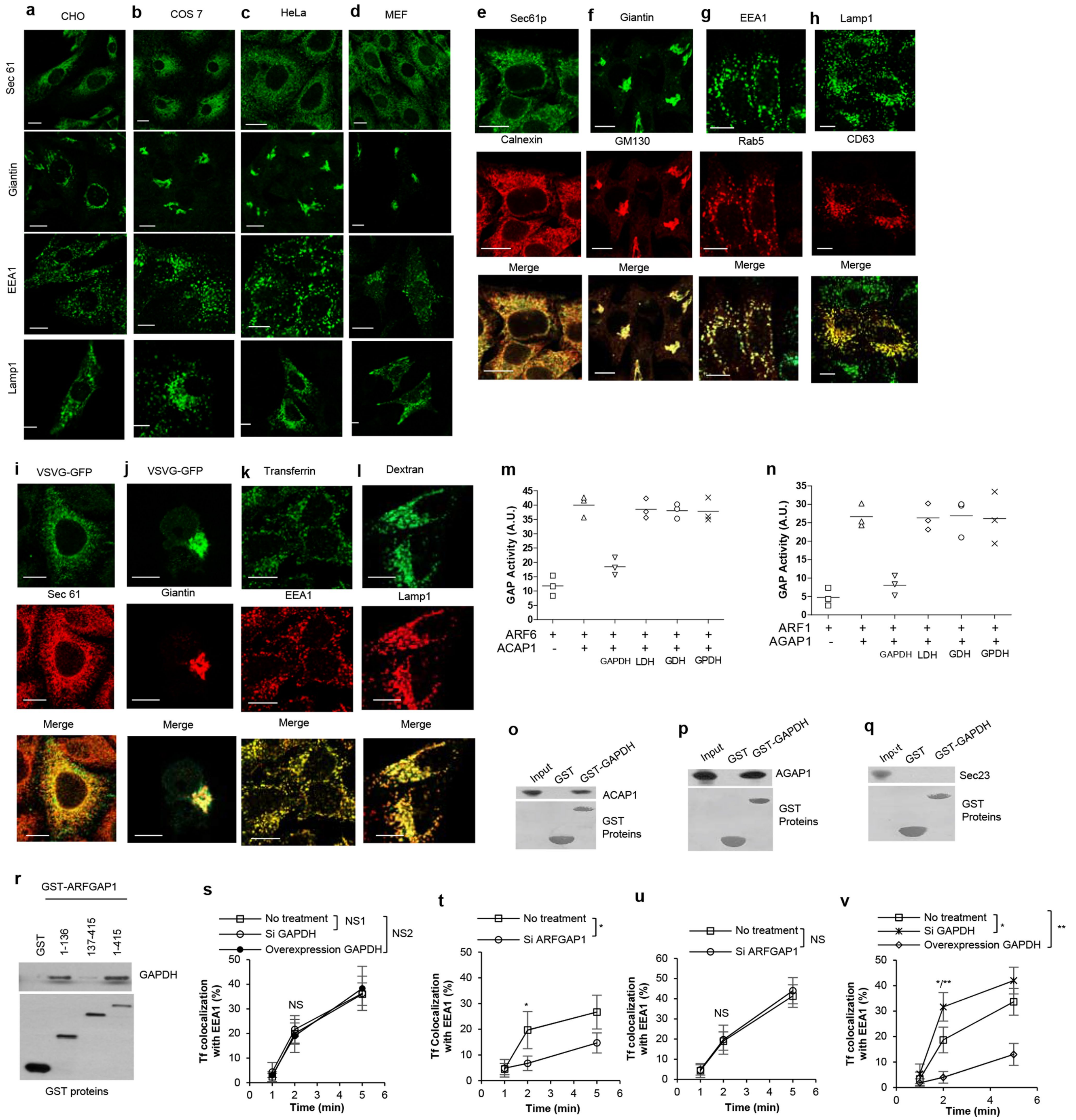
Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Further characterization of GAPDH inhibition of COPI transport. a, **b**, The COPI transport assay was performed on HeLa cells that were treated as indicated. A confocal image from a representative experiment (out of three) is shown. Green, VSVG–KDELR; red, giantin. Scale bar, 10 μm. Line-scan analysis for each image is also shown. **c**, The COPI transport assay was performed on HeLa cells that were treated as indicated. Quantification of a representative experiment (out of three) is shown, $n = 10$ fields of cells examined. Data are mean ± s.d. $^{*}P = 2.8 \times 10^{-7}$, $^{**}P = 8.2 \times 10^{-7}$. **d**, HeLa cells were treated as indicated, and cell death was quantified by flow cytometry. Data are mean ± s.d. from $n = 10$ independent experiments. $P = 0.7314$ (NS). **e**, HeLa cells were treated as indicated, and the COPI transport assay was performed. A confocal image from a representative experiment (out of three) is shown. Green, VSVG–KDELR; red, giantin. Scale bar, 10 μm. Line-scan analysis for each image is also shown. **f**, HeLa cells were treated as indicated, and whole-cell lysates were immunoblotted for proteins as indicated. Representative blot from three independent experiments. **g**, GST fusion proteins were incubated with purified GAPDH in a pull-down experiment, followed by immunoblotting as indicated. Representative blot from three independent experiments. **h**, HeLa cells were transfected with constructs as indicated, followed by immunoprecipitation using the Myc tag and then immunoblotting for GAPDH. Representative blot from two independent experiments. **i**, GST fusion proteins as indicated were bound to beads and then incubated with purified coatomer in a pull-down experiment, followed by immunoblotting to detect β-COP or Coomassie staining to detect GST fusion proteins. Representative results from three independent experiments. **j**, The indicated cytoplasmic tails of cargo proteins were fused to GST, bound to beads, and then incubated with ARFGAP1, followed by immunoblotting with antibody against ARFGAP1 or Coomassie staining to detect GST fusion proteins. Representative results from two independent experiments. **k**, HeLa cells were treated as indicated, and assayed for COPI transport. Quantification of a representative experiment (of three) is shown. Data are mean ± s.d. from $n = 10$ fields of cells examined. $^{*}P = 1.4 \times 10^{-6}$. **l**, HeLa cells were transfected with GAPDH constructs as indicated. ARFGAP1 immunoprecipitates were immunoblotted for the different forms of GAPDH as indicated. Representative blot from two experiments. *P* values determined by two-tailed *t*-test.

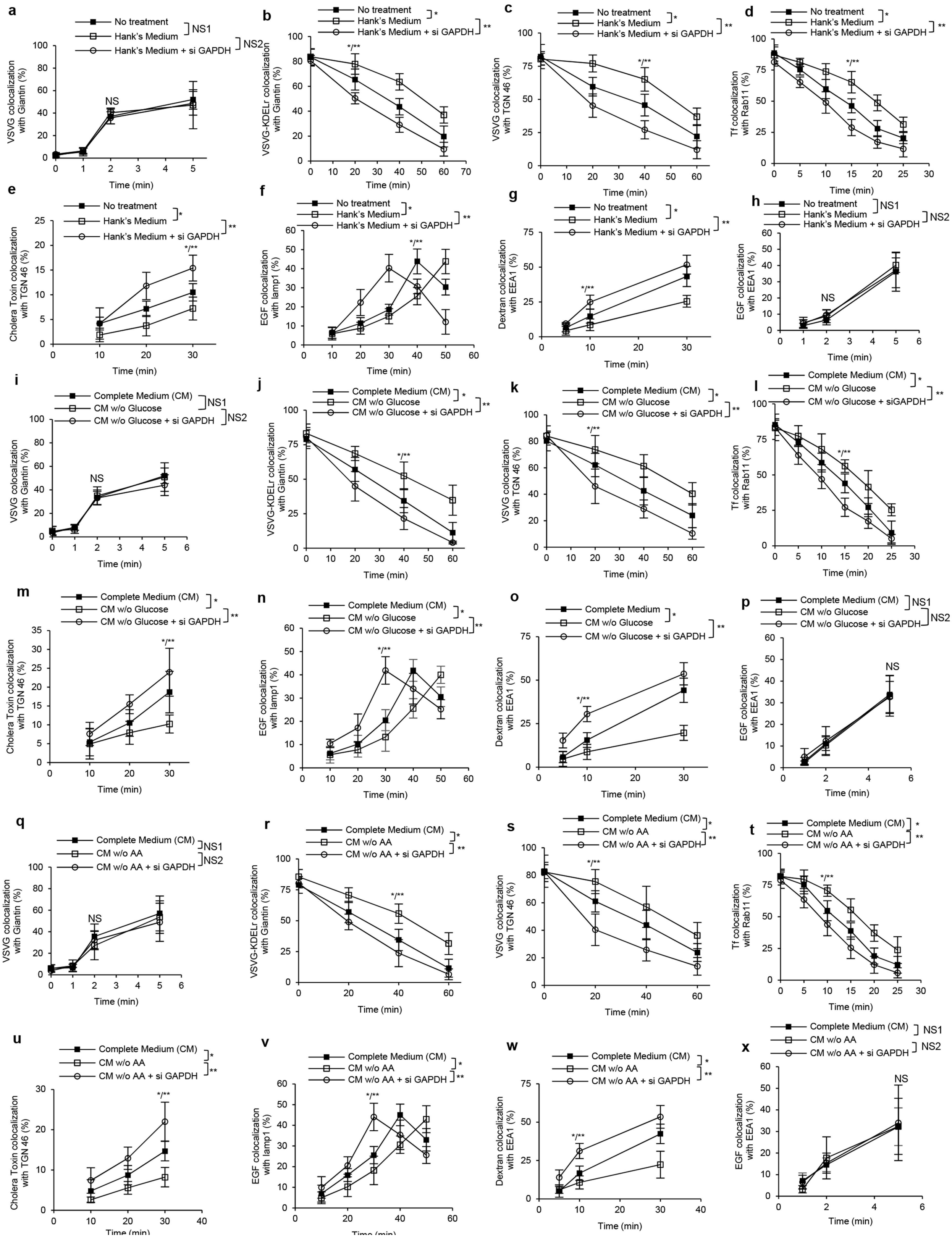

a
Control
Starvation
C
M
C
M
GAPDH
Actin
Cellubrevin
b
Control
Starvation
Giantin / GAPDH
Colocalization of GAPDH with Giantin (%)
*
Control
Starvation
c
Control
Starvation
TGN46 / GAPDH
Colocalization of GAPDH with TGN 46 (%)
*
d
Control
Starvation
EEA1 / GAPDH
Colocalization of GAPDH with EEA1 (%)
*
e
Control
Starvation
Lamp1 / GAPDH
Colocalization of GAPDH with Lamp1 (%)
*
f
Control
Starvation
Sec61 / GAPDH
Colocalization of GAPDH with Sec61 (%)
NS
g
5 min
Control
Si GAPDH
VSVG / Giantin
Intensity (Arbi. Unit)
Distance (Pixels)
h
20 min
Control
Si GAPDH
VSVG / TGN-46
i
10 min
Control
Si GAPDH
Rab11/ Transferrin
j
30 min
Control
Si GAPDH
TGN46 / CT
k
30 min
Control
Si GAPDH
Lamp1/ EGF
l
10 min
Control
Si GAPDH
EEA1/ Dextran
m
No treatment
Si GAPDH
Overexpression GAPDH
NS1
NS2
Cholera Toxin Colocalization with EEA1 (%)
NS
Time (min)
n
No treatment
Si GAPDH
Overexpression GAPDH
NS1
NS2
IL 2R Colocalization with EEA1 (%)
NS
Time (min)
o
R α GAPDH (Sigma)
R α GAPDH (Santa Cruz)
M α GAPDH (Ambion)
Control
p
GFP-GAPDH
q
si GAPDH
r
Rescue

Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | Further characterizing how GAPDH affects other pathways. a, HeLa cells were treated as indicated, fractionated into cytosol (C) and total membrane (M), and immunoblotted for proteins as indicated. Representative blot from two independent experiments. **b**–**f**, HeLa cells were treated as indicated, and examined for the colocalization between GAPDH and different organelle markers. A confocal image from a representative experiment (out of three) is shown. Scale bar, 10 μm (left); quantification of colocalization (right). Data are mean ± s.d. of $n = 10$ fields of cells examined. Colocalization of GAPDH (red) with giantin (green), $^*P = 2.9 \times 10^{-4}$ (**b**); colocalization of GAPDH (red) with TGN46 (green), $^*P = 1.5 \times 10^{-4}$ (**c**); colocalization of GAPDH (red) with EEA1 (green), $^*P = 1.9 \times 10^{-3}$ (**d**); colocalization of GAPDH (red) with LAMP1 (green), $^*P = 7.2 \times 10^{-4}$ (**e**); colocalization of GAPDH (red) with Sec61p (green), $P = 0.8531$ (**f**). **g**–**l**, Transport assays were performed in HeLa cells. A confocal image from a representative experiment (out of three) is shown. Scale bar, 10 μm. Line-scan analysis for the representative image is also shown. Transport from the ER to the Golgi (**g**; green, VSVG; red, giantin), transport from the Golgi to the plasma membrane (**h**; green, VSVG; red, TGN46), transport from the early endosome to the plasma membrane (**i**; red, transferrin; green, RAB11), transport from the plasma membrane to the Golgi (**j**; red, cholera toxin; green, TGN46); transport from the plasma membrane to the lysosome (**k**; red, EGF; green, LAMP1), transport from the plasma membrane to the early endosome (**l**; red, dextran; green, EEA1). **m**, **n**, Transport assays in HeLa cells; a representative experiment (out of three) is shown, $n = 10$ fields of cells examined. Data are mean ± s.d. Cholera toxin endocytosis (**m**), $P = 1$ (NS1); $P = 0.3205$ (NS2); interleukin-2 receptor beta subunit (IL2R-β) endocytosis (**n**), $P = 0.063$ (NS1), $P = 0.9264$ (NS2). **o**–**r**, HeLa cells were treated as indicated followed by immunofluorescence microscopy using different antibodies against GAPDH (**o**, **q**) or imaging of GFP-tagged GAPDH (**p**, **r**). Image from a representative experiment of two is shown. Scale bar, 10 μm. P values determined by two-tailed t-test.

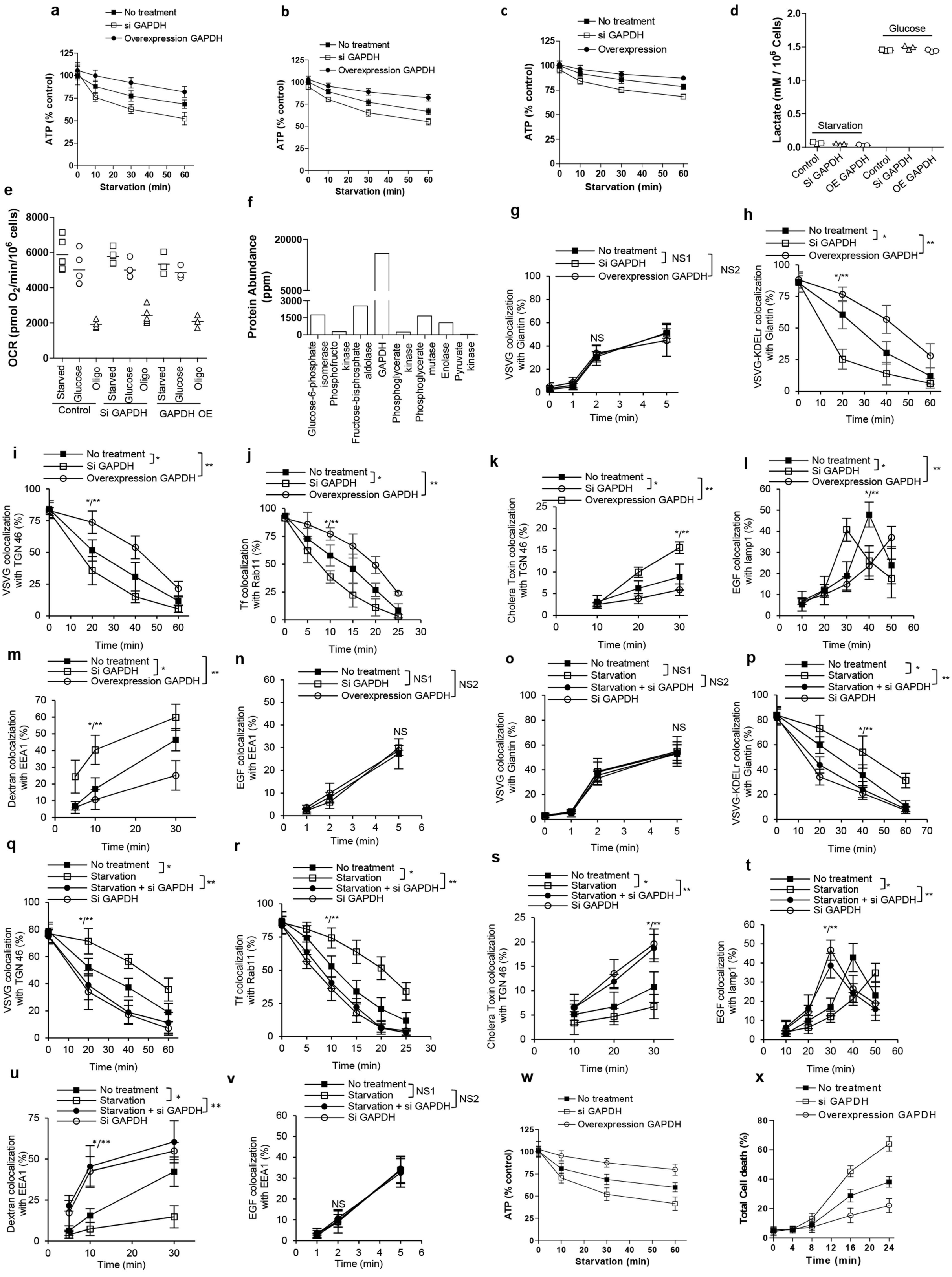

a CHO
b COS 7
c HeLa
d MEF
Sec 61
Giantin
EEA1
Lamp1
e Sec61p
f Giantin
g EEA1
h Lamp1
Calnexin
GM130
Rab5
CD63
Merge
i VSVG-GFP
j VSVG-GFP
k Transferrin
l Dextran
Sec 61
Giantin
EEA1
Lamp1
Merge
m
GAP Activity (A.U.)
ARF6
ACAP1
GAPDH
LDH
GDH
GPDH
n
GAP Activity (A.U.)
ARF1
AGAP1
GAPDH
LDH
GDH
GPDH
o
Input
GST
GST-GAPDH
ACAP1
GST Proteins
p
Input
GST
GST-GAPDH
AGAP1
GST Proteins
q
Input
GST
GST-GAPDH
Sec23
GST Proteins
r
GST-ARFGAP1
GST
1-136
137-415
1-415
GAPDH
GST proteins
s
No treatment
Si GAPDH
Overexpression GAPDH
NS1
NS2
NS
Tf colocalization with EEA1 (%)
Time (min)
t
No treatment
Si ARFGAP1
*
Tf Colocalization with EEA1 (%)
Time (min)
u
No treatment
Si ARFGAP1
NS
Tf colocalization with EEA1 (%)
Time (min)
v
No treatment
Si GAPDH
Overexpression GAPDH
*
**
*/**
Tf colocalization with EEA1 (%)
Time (min)

Extended Data Fig. 3 | Additional characterization of GAPDH and its effects on the transport pathways. a–d, The distribution of different organelle markers in CHO (**a**), COS-7 (**b**), HeLa (**c**) or MEF (**d**) cells were assessed by immunofluorescence microscopy. Image from a representative experiment of two experiments is shown. Scale bar, 10 μm. **e–h**, Comparison of the distribution of two markers against the same intracellular compartment using confocal microscopy. ER markers Sec61p (green) and calnexin (red) (**e**); Golgi markers giantin (green) and GM130 (red) (**f**); early endosome markers EEA1 (green) and RAB5 (red) (**g**); lysosome markers LAMP1 (green) and CD63 (red) (**h**). Images are representative of two experiments. Scale bar, 10 μm. **i–l**, Confirming the staining specificity of organelle markers using model cargoes that reside at specific intracellular compartments using confocal microcopy: GFP-tagged VSVG at the ER (green) and Sec61p (red) (**i**); GFP-tagged VSVG at the Golgi (green) and giantin (red) (**j**); fluorescently labelled transferrin (green) and EEA (red) (**k**); fluorescently labelled dextran (green) and LAMP1 (red) (**l**) Images are representative of two experiments. Scale bar, 10 μm. **m**, **n**, The GAP assay was performed using ARF6 and ACAP1 (**m**) or ARF1 and AGAP1 (**n**), in the presence of different metabolic enzymes as indicated. One experiment of three independent experiments is shown. **o–r**, Pull-down studies to detect GAPDH binding directly to: ACAP1 (**o**), AGAP1 (**p**), Sec23p (**q**) or different portions of ARFGAP1 as indicated (**r**). Representative blot from three experiments. **s–v**, HeLa (**s**, **u**) or BSC-1 (**t**, **v**) cells were treated as indicated, followed by transport assay for transferrin endocytosis. Quantification of a representative experiment (out of three) is shown, $n = 10$ fields of cells examined. Data are mean ± s.d. In **s**: $P = 0.4646$ (NS1); $P = 0.6973$ (NS2). In **t**: $*P = 1.8 \times 10^{-4}$. In **u**: $P = 0.8073$ (NS). In **v**: $*P = 3.6 \times 10^{-5}$, $**P = 1.7 \times 10^{-6}$. P values determined by two-tailed t-test.



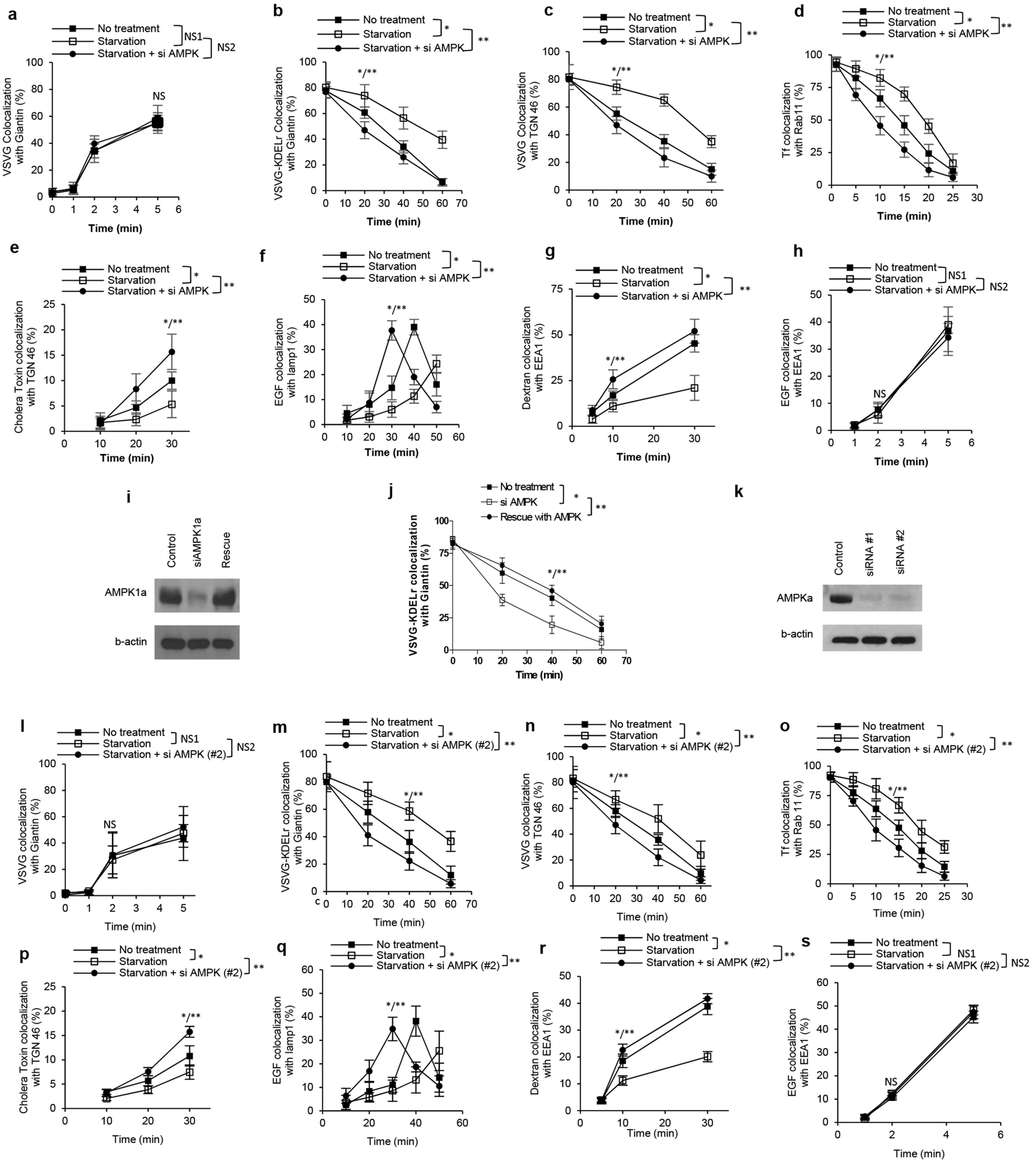
Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | Different ways of starving cells lead to identical pathways being inhibited and inhibition requires GAPDH. **a**–**h**, Transport assays were performed following incubation of HeLa cells in Hank's medium. Quantification of one experiment (out of three) is shown. Data are mean ± s.d. of $n = 10$ fields of cells examined. Transport of VSVG from the ER to the Golgi, $P = 0.1944$ (NS1), $P = 0.05$ (NS2) (**a**); COPI transport of VSVG–KDELR from the Golgi to the ER, *$P = 2.4 \times 10^{-6}$, **$P = 4.2 \times 10^{-10}$ (**b**); transport of VSVG from the Golgi to the plasma membrane, *$P = 1.7 \times 10^{-5}$, **$P = 6.7 \times 10^{-8}$ (**c**); transport of transferrin from the early endosome to the plasma membrane, *$P = 2.9 \times 10^{-5}$, **$P = 6.9 \times 10^{-9}$ (**d**); transport of cholera toxin from the plasma membrane to the Golgi, *$P = 2.4 \times 10^{-3}$, **$P = 7.8 \times$ `10^{-7} (**e**); transport of EGF from the plasma membrane to the lysosome, *$P = 1.9 \times 10^{-6}$, **$P = 1.5 \times 10^{-2}$ (**f**); transport of dextran from the plasma membrane to the early endosome, *$P = 7.8 \times 10^{-3}$, **$P = 3.4 \times \times 10^{-7}$ (**g**); transport of EGF from the plasma membrane to the early endosome, $P = 0.1485$ (NS1), $P = 0.6378$ (NS2) (**h**). **i**–**p**, HeLa cells were incubated in medium without glucose, and then transport assays were performed. Quantification of one experiment (out of three) is shown. Data are mean ± s.d. from $n = 10$ fields of cells examined. Transport of VSVG from the ER to the Golgi, $P = 0.6921$ (NS1), $P = 0.5648$ (NS2) (**i**); COPI transport of VSVG–KDELR from the Golgi to the ER, *$P = 4.4 \times 10^{-4}$, **$P = 6.5 \times 10^{-7}$ (**j**); transport of VSVG from the Golgi to the plasma membrane, *$P = 1.6 \times 10^{-2}$, **$P = 6.7 \times 10^{-5}$ (**k**); transport of transferrin from the early endosome to the plasma membrane, *$P = 1.9 \times 10^{-4}$, **$P = 2.8 \times 10^{-9}$ (**l**); transport of cholera toxin from the plasma membrane to the Golgi, *$P = 7.4 \times 10^{-4}$, **$P = 4.2 \times 10^{-5}$ (**m**); transport of EGF from the plasma membrane to the lysosome, *$P = 7.6 \times 10^{-3}$, **$P = 3.2 \times 10^{-9}$ (**n**); transport of dextran from the plasma membrane to the early endosome, *$P = 2.2 \times 10^{-3}$, **$P = 1.6 \times 10^{-5}$ (**o**); transport of EGF from the plasma membrane to the early endosome, $P = 0.8648$ (NS1), $P = 0.8946$ (NS2) (**p**). **q**–**x**, HeLa cells were incubated in medium without amino acids, and transport assays were performed. Quantification of a representative experiment (out of three) is shown. Data are mean ± s.d. of $n = 10$ fields of cells examined. Transport of VSVG from the ER to the Golgi, $P = 0.1419$ (NS1), $P = 0.3379$ (NS2) (**q**); COPI transport of VSVG–KDELR from the Golgi to the ER, *$P = 1.9 \times 10^{-5}$, **$P = 1.1 \times 10^{-6}$ (**r**); transport of VSVG from the Golgi to the plasma membrane, *$P = 1.1 \times 10^{-3}$, **$P = 6.6 \times 10^{-7}$ (**s**); transport of transferrin from the early endosome to the plasma membrane, *$P = 7.2 \times 10^{-5}$, **$P = 8.4 \times 10^{-7}$ (**t**); transport of cholera toxin from the plasma membrane to the Golgi, *$P = 1.1 \times 10^{-5}$, **$P = 1.8 \times 10^{-6}$ (**u**); transport of EGF from the plasma membrane to the lysosome, *$P = 1.1 \times 10^{-2}$, **$P = 9.3 \times 10^{-8}$ (**v**); transport of dextran from the plasma membrane to the early endosome, *$P = 7.8 \times 10^{-3}$, **$P = 1.2 \times 10^{-8}$ (**w**); transport of EGF from the plasma membrane to the early endosome, $P = 0.9731$ (NS1), $P = 0.8159$ (NS2) (**x**). P values determined by two-tailed t-test.



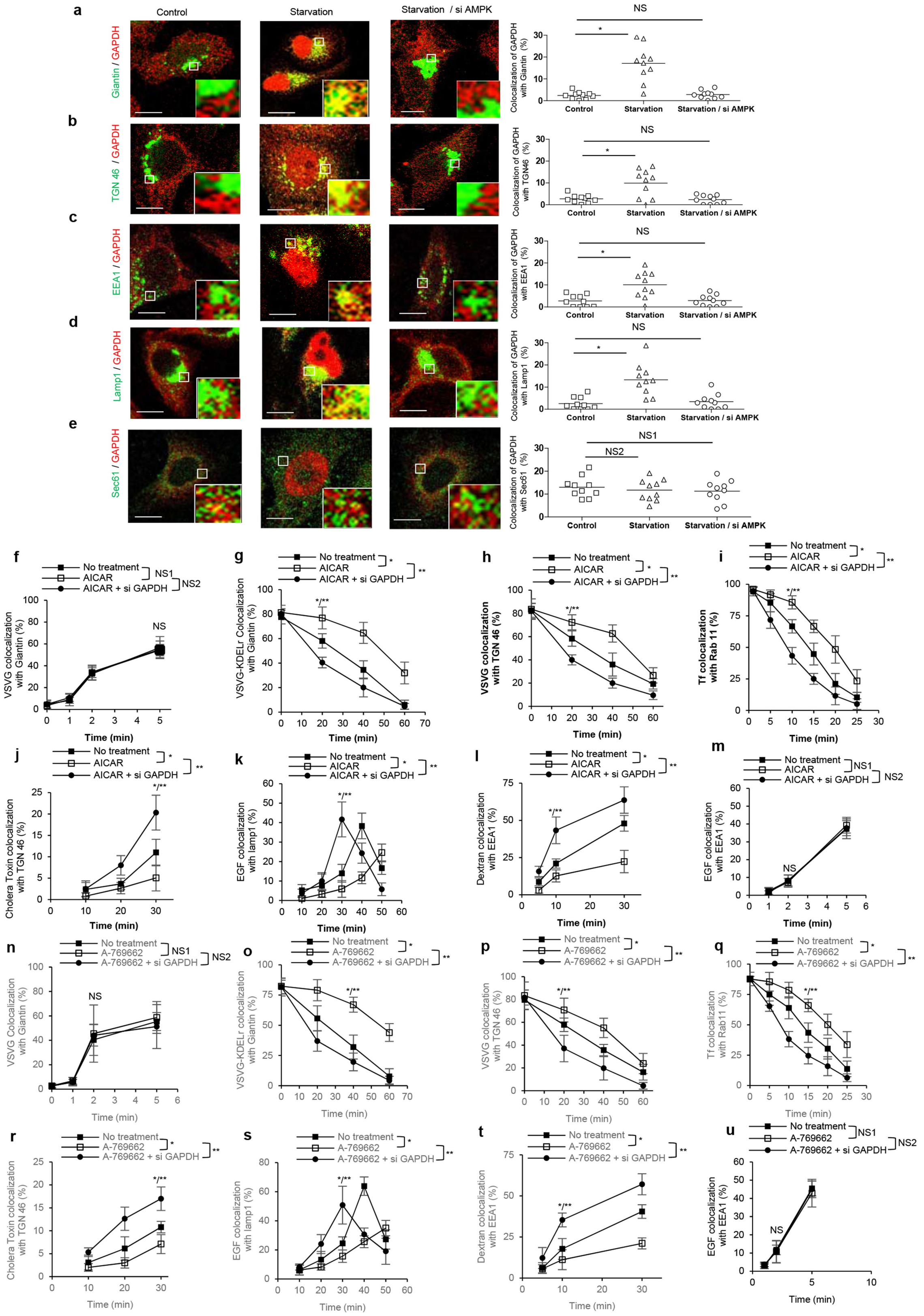
Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | Further characterizing the effects of starvation and GAPDH. a–c, HeLa cells were starved using Hank's medium (**a**), medium lacking glucose (**b**), or medium lacking amino acids (**c**), and the total ATP level was measured. Results of one experiment representative of three independent experiments are shown. **d**, Lactate production, measured in HeLa cells treated as indicated. Results of one experiment representative of three independent experiments are shown. **e**, Oxygen consumption rate in HeLa cells treated as indicated. Results of one experiment representative of three independent experiments is shown. **f**, Relative abundance of glycolytic enzymes in HeLa cells. Data are derived from https://pax-db.org; glucose-6-phosphate isomerase (https://pax-db.org/protein/1860535); phosphofructokinase (https://pax-db.org/protein/1852972); fructose-bisphosphate aldolase (https://pax-db.org/protein/1851945); GAPDH (https://pax-db.org/protein/1843185); phosphoglycerate kinase (https://pax-db.org/protein/1848799); phosphoglycerate mutase (https://pax-db.org/protein/1855378); enolase (https://pax-db.org/protein/1850612); pyruvate kinase (https://pax-db.org/protein/1852273). **g–n**, Transport assays were performed on HEK293 cells that were treated as indicated. Quantification of one representative experiment (out of three) is shown. Data are mean ± s.d. from $n = 10$ fields of cells examined. Transport of VSVG from the ER to the Golgi, $P = 0.6468$ (NS1), $P = 0.2133$ (NS2) (**g**); COPI transport of VSVG–KDELR from the Golgi to the ER, $*P = 4.3 \times 10^{-7}$, $**P = 1.3 \times 10^{-3}$ (**h**); transport of VSVG from the Golgi to the plasma membrane, $*P = 2.1 \times 10^{-3}$, $**P = 1.9 \times 10^{-5}$ (**i**); transport of transferrin from the early endosome to the plasma membrane, $*P = 6.3 \times 10^{-5}$, $**P = 4.7 \times 10^{-4}$ (**j**); transport of cholera toxin from the plasma membrane to the Golgi, $*P = 6.5 \times 10^{-4}$, $**P = 1.5 \times 10^{-2}$ (**k**); transport of EGF from the plasma membrane to the lysosome, $*P = 7.8 \text{ x } 10^{-7}$, $**P = 6.5 \times 10^{-8}$ (**l**); transport of dextran from the plasma membrane to the early endosome, $*P = 3.6 \times 10^{-6}$, $**P = 3.4 \times 10^{-2}$ (**m**); transport of EGF from the plasma membrane to the early endosome, $P = 0.2375$ (NS1), $P = 0.5291$ (NS2) (**n**). **o–v**, Transport assays were performed on HEK293 cells that were treated as indicated. Quantification of a representative experiment (out of three) is shown. Data are mean ± s.d. from $n = 10$ fields of cells examined. Transport of VSVG from the ER to the Golgi, $P = 0.6649$ (NS1), $P = 0.84$ (NS2) (**o**); COPI transport of VSVG–KDELR from the Golgi to the ER, $*P = 4.6 \times 10^{-3}$, $**P = 2.5 \times 10^{-6}$ (**p**); transport of VSVG from the Golgi to the plasma membrane, $*P = 7.1 \times 10^{-5}$, $**P = 1.1 \times 10^{-6}$ (**q**); transport of transferrin from the early endosome to the plasma membrane, $*P = 7.2 \times 10^{-6}$, $**P = 6.5 \times 10^{-9}$ (**r**); transport of cholera toxin from the plasma membrane to the Golgi, $*P = 5.6 \times 10^{-3}$, $**P = 8.8 \times 10^{-9}$ (**s**) transport of EGF from the plasma membrane to the lysosome, $*P = 1.1 \times 10^{-2}$, $**P = 2.9 \times 10^{-8}$ (**t**); transport of dextran from the plasma membrane to the early endosome, $*P = 2.4 \times 10^{-4}$, $**P = 2.4 \times 10^{-6}$ (**u**); transport of EGF from the plasma membrane to the early endosome, $P = 0.7924$ (NS1), $P = 0.4675$ (NS2) (**v**). **w**, **x**, HEK293 cells were treated as indicated, followed by starvation and quantification of total ATP level (**w**) or cell death (**x**). Data are mean ± s.d. of $n = 3$ experiments. P values determined by two-tailed t-test.



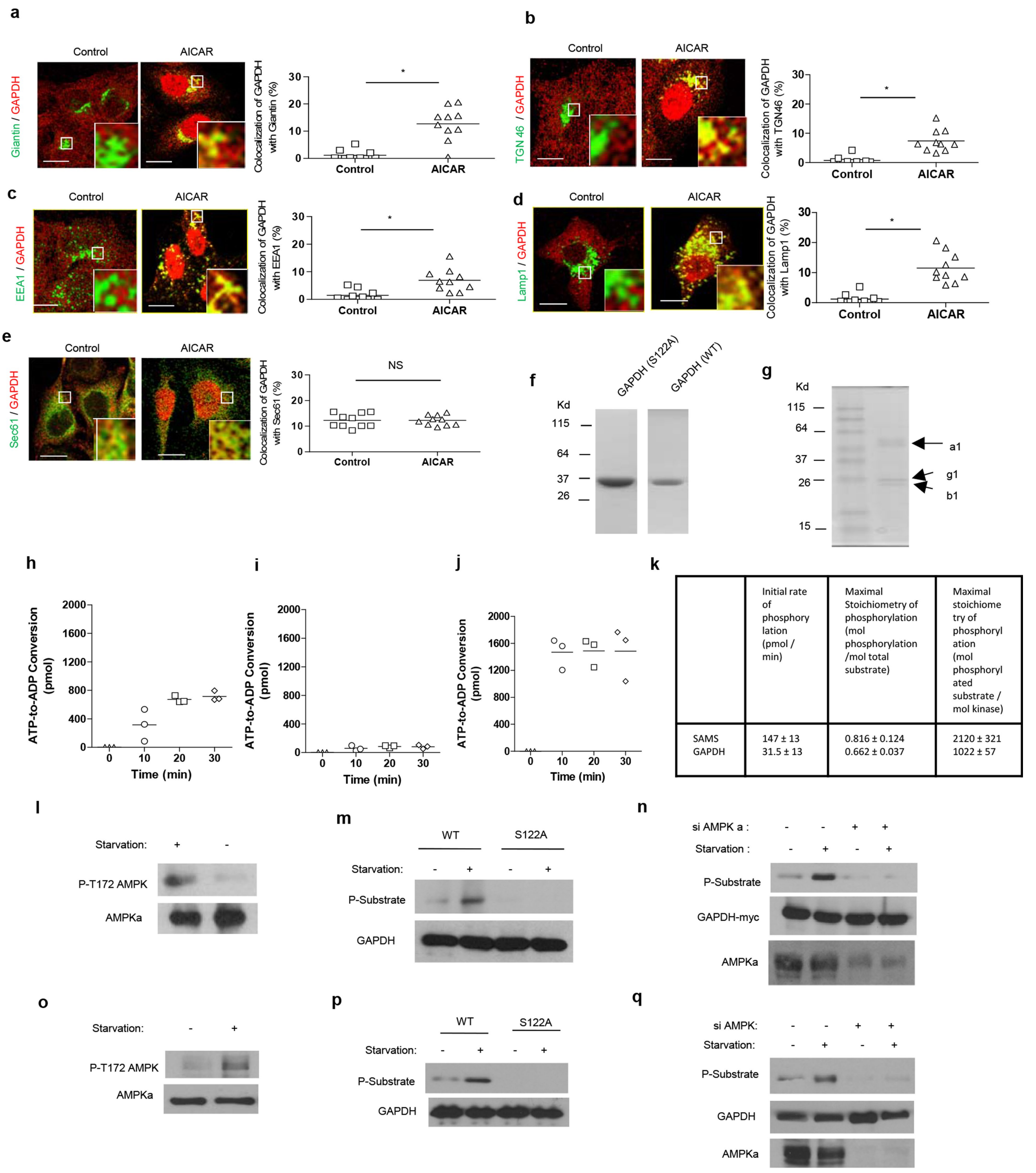
Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | Inhibition of transport pathways by starvation and AMPK. a–h, HeLa cells were starved using a general starvation medium (lacking glucose and amino acids), and transport assays were performed. Quantification of a representative experiment (out of three) is shown. Data are mean ± s.d. from $n = 10$ fields of cells examined. Transport of VSVG from the ER to the Golgi, $P = 0.3506$ (NS1), $P = 0.9126$ (NS2) (**a**); COPI transport of VSVG–KDELR from the Golgi to the ER, $^{*}P = 8.6 \times 10^{-4}$, $^{**}P = 5.8 \times 10^{-7}$ (**b**); transport of VSVG from the Golgi to the plasma membrane, $^{*}P = 1.0 \times 10^{-7}$, $^{**}P = 4.4 \times 10^{-9}$ (**c**); transport of transferrin from the early endosome to the plasma membrane, $^{*}P = 4.2 \times 10^{-5}$, $^{**}P = 4.4 \times 10^{-10}$ (**d**); transport of cholera toxin from the plasma membrane to the Golgi, $^{*}P = 2.6 \times 10^{-4}$, $^{**}P = 9.5 \times 10^{-7}$ (**e**); transport of EGF from the plasma membrane to the lysosome, $^{*}P = 2.0 \times 10^{-4}$, $^{**}P = 2.2 \times 10^{-13}$ (**f**); transport of dextran from the plasma membrane to the early endosome, $^{*}P = 1.4 \times 10^{-4}$, $^{**}P = 1.6 \times 10^{-6}$ (**g**); transport of EGF from the plasma membrane to the early endosome, $P = 0.1085$ (NS1), $P = 0.1408$ (NS2) (**h**). **i**, HeLa cells were treated as indicated and cell lysate was immunoblotted for proteins as indicated. Blot is representative of two independent experiments. **j**, HeLa cells were starved, and the COPI transport assay was performed. Quantification of a representative experiment (out of three) is shown. Data are mean ± s.d. of $n = 10$ fields of cells examined. $^{*}P = 1.1 \times 10^{-4}$, $^{**}P = 3.5 \times 10^{-6}$. **k**, HeLa cells were treated as indicated and the whole-cell lysate was immunoblotted for proteins as indicated. Representative blot of two independent experiments. **l–s**, HeLa cells were starved, and transport assays were performed. Quantification of a representative experiment (out of three) is shown. Data are mean ± s.d. from $n = 10$ fields of cells examined. Transport of VSVG from the ER to the Golgi, $P = 0.6077$ (NS1), $P = 0.5535$ (NS2) (**l**); COPI transport of VSVG–KDELR from the Golgi to the ER, $^{*}P = 3.5 \times 10^{-6}$, $^{**}P = 3.6 \times 10^{-10}$ (**m**); transport of VSVG from the Golgi to the plasma membrane, $^{*}P = 4.1 \times 10^{-3}$, $^{**}P = 4.4 \times 10^{-6}$ (**n**); transport of transferrin from the early endosome to the plasma membrane, $^{*}P = 3.7 \times 10^{-6}$, $^{**}P = 1.1 \times 10^{-9}$ (**o**); transport of cholera toxin from the plasma membrane to the Golgi, $^{*}P = 7.6 \times 10^{-4}$, $^{**}P = 8.7 \times 10^{-11}$ (**p**) transport of EGF from the plasma membrane to the lysosome, $^{*}P = 3.1 \times 10^{-8}$, $^{**}P = 1.3 \times 10^{-2}$ (**q**) transport of dextran from the plasma membrane to the early endosome, $^{*}P = 1.1 \times 10^{-6}$, $^{**}P = 3.7 \times 10^{-10}$ (**r**) transport of EGF from the plasma membrane to the early endosome, $P = 0.228$ (NS1), $P = 0.1738$ (NS2) (**s**). *P* values determined by two-tailed *t*-test.



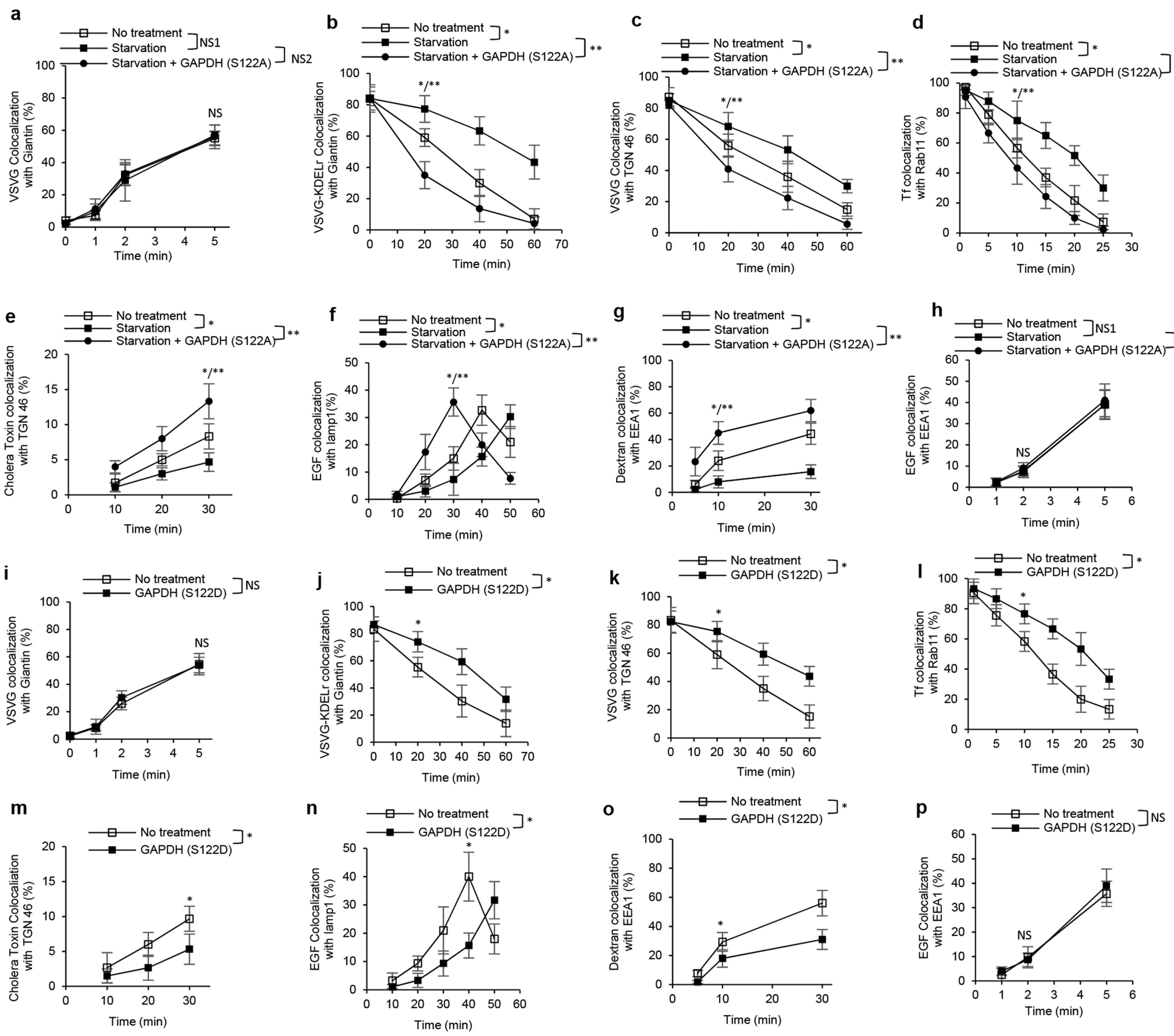
Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Effects of AMPK on GAPDH distribution and transport pathways. **a**–**e**, HeLa cells were treated as indicated, and then examined for colocalization between GAPDH and different organelle markers. A confocal image from a representative experiment (out of three) is shown. Scale bar, 10 μm (left). quantification is also shown (right). Data are mean ± s.d. of $n = 10$ fields of cells examined. Colocalization of GAPDH (red) with giantin (green), $*P = 2.9 \times 10^{-3}$, $P = 0.6397$ (NS) (**a**); colocalization of GAPDH (red) with TGN46 (green), $*P = 6.5 \times 10^{-3}$, $P = 0.6413$ (NS) (**b**); colocalization of GAPDH (red) with EEA1 (green), $*P = 3.5 \times 10^{-3}$, $P = 0.8793$ (NS) (**c**); colocalization of GAPDH (red) with LAMP1 (green), $*P = 9.4 \times 10^{-4}$, $P = 0.551$ (NS) (**d**); colocalization of GAPDH (red) with Sec61p (green), $P = 0.5361$ (NS1), $P = 0.4243$ (NS2) (**e**). **f**–**m**, HeLa cells were treated as indicated, and transport assays were performed. Quantification of a representative experiment (out of three) is shown. Data are mean ± s.d. from $n = 10$ fields of cells examined. Transport of VSVG from the ER to the Golgi, $P = 0.5250$ (NS1), $P = 0.8291$ (NS2) (**f**); COPI transport of VSVG–KDELR from the Golgi to the ER, $*P = 4.6 \times 10^{-5}$, $**P = 2.9 \times 10^{-8}$ (**g**); transport of VSVG from the Golgi to the plasma membrane, $*P = 1.4 \times 10^{-4}$, $**P = 6.9 \times 10^{-10}$ (**h**); transport of transferrin from the early endosome to the plasma membrane, $*P = 2.0 \times 10^{-7}$, $**P = 1.0 \times 10^{-11}$ (**i**) transport of cholera toxin from the plasma membrane to the Golgi, $*P = 4.0 \times 10^{-4}$, $**P = 3.9 \times 10^{-8}$ (**j**); transport of EGF from the plasma membrane to the lysosome, $*P = 7.1 \times 10^{-4}$, $**P = 4.2 \times 10^{-8}$ (**k**); transport of dextran from the plasma membrane to the early endosome, $*P = 1.4 \times 10^{-7}$, $**P = 1.9 \times 10^{-9}$ (**l**); transport of EGF from the plasma membrane to the early endosome, $P = 0.6088$ (NS1), $P = 0.4061$ (NS2) (**m**). **n**–**u**, HeLa cells were treated as indicated, and transport assays were performed. Quantification of a representative experiment (out of three) is shown. Data are mean ± s.d. from $n = 10$ fields of cells examined. Transport of VSVG from the ER to the Golgi, $P = 0.5559$ (NS1), $P = 0.8576$ (NS2) (**n**); COPI transport of VSVG–KDELR from the Golgi to the ER, $*P = 4.1 \times 10^{-5}$, $**P = 2.3 \times 10^{-9}$ (**o**); transport of VSVG from the Golgi to the plasma membrane, $*P = 4.2 \times 10^{-3}$, $**P = 6.3 \times 10^{-5}$ (**p**); transport of transferrin from the early endosome to the plasma membrane, $*P = 7.2 \times 10^{-4}$, $**P = 3.7 \times 10^{-11}$ (**q**) transport of cholera toxin from the plasma membrane to the Golgi, $*P = 2.3 \times 10^{-4}$, $**P = 3.0 \times 10^{-8}$ (**r**); transport of EGF from the plasma membrane to the lysosome, $*P = 2.9 \times 10^{-3}$, $**P = 5.8 \times 10^{-6}$ (**s**), transport of dextran from the plasma membrane to the early endosome, $*P = 2.9 \times 10^{-2}$, $**P = 1.6 \times 10^{-8}$ (**t**) transport of EGF from the plasma membrane to the early endosome, $P = 0.6856$ (NS1), $P = 0.9248$ (NS2) (**u**). P values determined by two-tailed t-test.

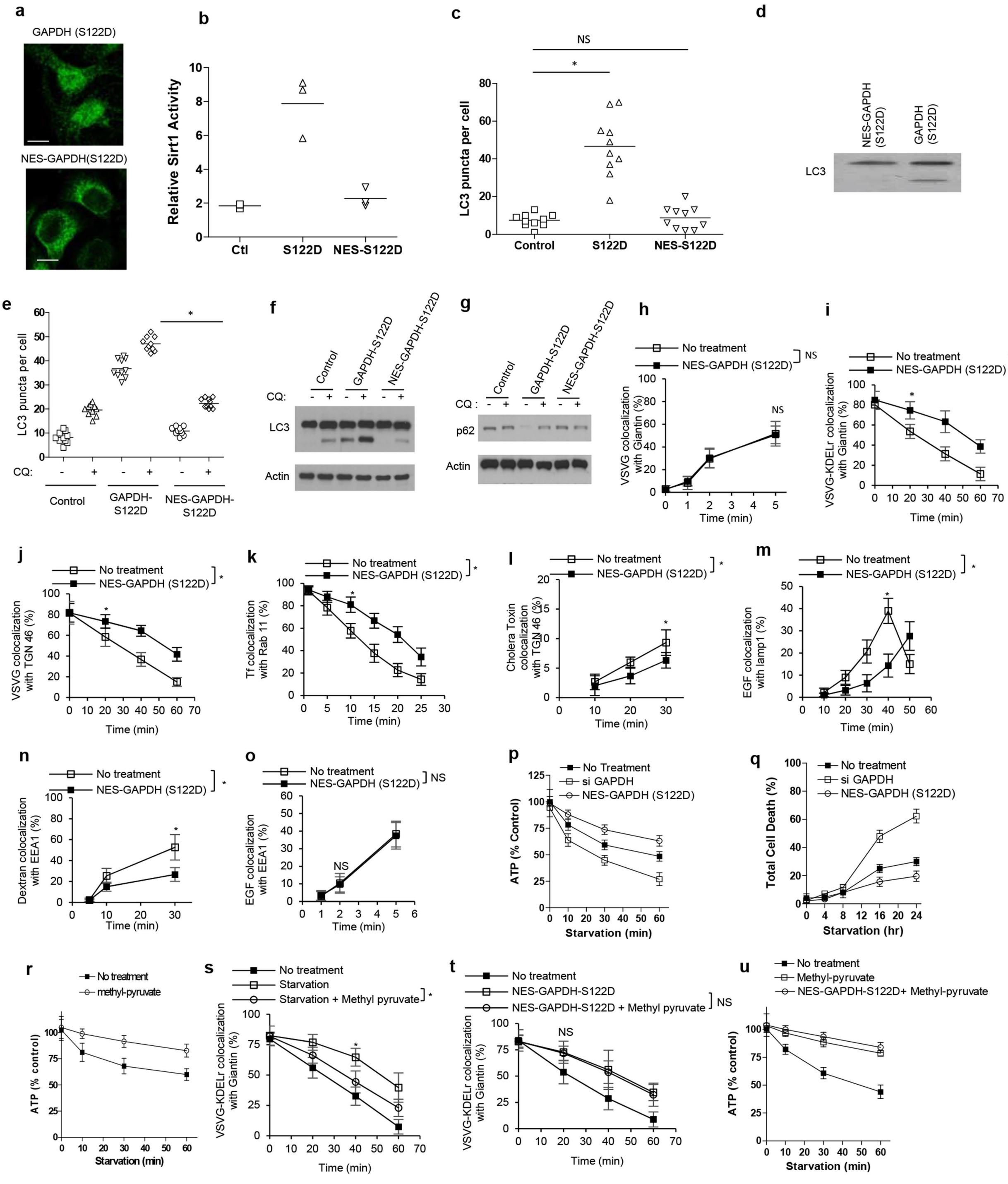


	Initial rate of phosphorylation (pmol / min)	Maximal Stoichiometry of phosphorylation (mol phosphorylation /mol total substrate)	Maximal stoichiometry of phosphorylated substrate / mol kinase)
SAMS GAPDH	147 ± 13 31.5 ± 13	0.816 ± 0.124 0.662 ± 0.037	2120 ± 321 1022 ± 57

Extended Data Fig. 8 | Further characterizing how AMPK acts on GAPDH. a–e, HeLa cells were treated as indicated, and then examined for colocalization between GAPDH and different organelle markers. A confocal image from a representative experiment (out of three) is shown (left) and quantified (right). Scale bar, 10 µm. Data are mean ± s.d. from $n = 10$ fields of cells examined. Colocalization of GAPDH (red) with giantin (green), $*P = 1.5 \times 10^{-4}$ (**a**); colocalization of GAPDH (red) with TGN46 (green), $*P = 2.2 \times 10^{-4}$ (**b**); colocalization of GAPDH (red) with EEA1 (green), $*P = 2.1 \times 10^{-3}$ (**c**); colocalization of GAPDH (red) with LAMP1 (green), $*P = 7.1 \times 10^{-5}$ (**d**); colocalization of GAPDH (red) with Sec61p (green), $P = 0.9933$ (NS) (**e**). **f**, Purity of GAPDH forms assessed by Coomassie gel staining. Representative gel from two independent experiments. **g**, Purity of AMPK complex assessed by Coomassie gel staining. Representative gel from two independent experiments. **h–j**, AMPK was incubated with wild-type GAPDH (**h**), GAPDH(S122A) (**i**), or SAMS, a peptide derived from acetyl-CoA carboxylase (**j**) in the in vitro kinase assay, and phosphorylation was measured over time. Data are mean ± s.d from a representative experiment of three independent experiments. **k**, Calculation of the stoichiometry of phosphorylation from the results shown in **h–j**. **l–q**, Whole-cell lysates, from HeLa (**l–n**) or HEK293 (**o–q**) cells, were treated as indicated and immunoblotted for proteins as indicated. Blots are representative of two independent experiments. *P* values determined by two-tailed *t*-test.



Extended Data Fig. 9 | Further characterizing how GAPDH mediates the ability of starvation to inhibit the transport pathways. **a–h**, HeLa cells were transfected with GAPDH(S122A), followed by starvation. A representative transport assay (of three experiments) is shown. Data are mean ± s.d. from $n = 10$ fields of cells examined. Transport of VSVG from the ER to the Golgi, $P = 0.4144$ (NS1), $P = 0.7463$ (NS2) (**a**); COPI transport of VSVG–KDELR from the Golgi to the ER, $*P = 3.8 \times 10^{-5}$, $**P = 1.9 \times 10^{-9}$ (**b**); transport of VSVG from the Golgi to the plasma membrane, $*P = 3.6 \times 10^{-3}$, $**P = 1.4 \times 10^{-6}$ (**c**); transport of transferrin from the early endosome to the plasma membrane, $*P = 1.4 \times 10^{-3}$, $**P = 1.4 \times 10^{-5}$ (**d**); transport of cholera toxin from the plasma membrane to the Golgi, $*P = 7.9 \times 10^{-5}$, $**P = 1.5 \times 10^{-7}$ (**e**); transport of EGF from the plasma membrane to the lysosome, $*P = 4.0 \times 10^{-3}$, $**P = 1.3 \times 10^{-9}$ (**f**); transport of dextran from the plasma membrane to the early endosome, $*P = 3.1 \times 10^{-5}$, $**P = 1.2 \times 10^{-8}$ (**g**); transport of EGF from the plasma membrane to the early endosome, $P = 0.7549$ (NS1), $P = 0.2279$ (NS2) (**h**). **i–p**, HeLa cells were transfected with GAPDH(S122D). A representative transport assay (of three experiments) is shown. Data are mean ± s.d. from $n = 10$ fields of cells examined. Transport of VSVG from the ER to the Golgi, $P = 0.8289$ (NS) (**i**); COPI transport of VSVG–KDELR from the Golgi to the ER, $*P = 2.2 \times 10^{-5}$ (**j**); transport of VSVG from the Golgi to the plasma membrane, $*P = 6.3 \times 10^{-4}$ (**k**); transport of transferrin from the early endosome to the plasma membrane, $*P = 6.9 \times 10^{-6}$ (**l**); transport of cholera toxin from the plasma membrane to the Golgi, $*P = 1.3 \times 10^{-4}$ (**m**); transport of EGF from the plasma membrane to the lysosome, $*P = 1.6 \times 10^{-3}$ (**n**); transport of dextran from the plasma membrane to the early endosome, $*P = 7.4 \times 10^{-4}$ (**o**); transport of EGF from the plasma membrane to the early endosome, $P = 0.5473$ (NS) (**p**). P values determined by two-tailed t-test.



Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | Further confirmation that the roles of GAPDH in transport and autophagy are distinct. a, HeLa cells were transfected with GFP-tagged forms of GAPDH as indicated and examined by immunofluorescence microscopy. A representative image from two independent experiments is shown. Scale bar, 10 µm. **b**, SIRT1 activity in HeLa cells transfected with GFP-tagged forms of GAPDH as indicated, Data are mean ± s.d. from three independent experiments. **c**, LC3 puncta formation in HeLa cells transfected with GFP-tagged forms of GAPDH as indicated. Quantification of a representative experiment (of three) is shown. Data are mean ± s.d. from $n = 10$ fields of cells examined. $*P = 2.3 \times 10^{-5}$, $P = 0.5598$ (NS). **d**, LC3 lipidation was assessed by immunoblotting of lysates from HeLa cells transfected with GFP-tagged forms of GAPDH as indicated. A representative blot from two independent experiments is shown. **e**, LC3 puncta formation in HeLa cells treated as indicated. Quantification of a representative experiment (of three) is shown. Data are mean ± s.d. from $n = 10$ fields of cells examined. $*P = 1.7 \times 10^{-5}$. **f**, LC3 lipidation was assessed by immunoblotting of lysates from HeLa cells treated as indicated. A representative blot from two independent experiments is shown. **g**, p62 level was assessed by immunoblotting of lysates from HeLa cells treated as indicated. A representative blot from two independent experiments is shown. **h**–**o**, Transport assays in HeLa cells treated as indicated. Quantification of a representative experiment (of three) is shown. Data are mean ± s.d. from $n = 10$ fields of cells examined. Transport of VSVG from the ER to the Golgi, $P = 0.8155$ (NS) (**h**); COPI transport of VSVG–KDELR from the Golgi to the ER, $*P = 1.4 \times 10^{-5}$ (**i**); transport of VSVG from the Golgi to the plasma membrane, $*P = 5.5 \times 10^{-4}$ (**j**); transport of transferrin from the early endosome to the plasma membrane, $*P = 3.3 \times 10^{-7}$ (**k**); transport of cholera toxin from the plasma membrane to the Golgi, $*P = 2.0 \times 10^{-3}$ (**l**); transport of EGF from the plasma membrane to the lysosome, $*P = 2.4 \times 10^{-6}$ (**m**); transport of dextran from the plasma membrane to the early endosome, $*P = 3.9 \times 10^{-5}$ (**n**); transport of EGF from the plasma membrane to the early endosome, $P = 0.7725$ (NS) (**o**). **p**, **q**, Total ATP level (**p**) and cell death (**q**) in HeLa cells treated as indicated. Data are mean ± s.d. from three independent experiments. **r**, Total ATP level in HeLa cells treated as indicated. Data are mean ± s.d. from three independent experiments. **s**, **t**, COPI transport in HeLa cells treated as indicated. Quantification of a representative experiment (of three) is shown. Data are mean ± s.d. from $n = 10$ fields of cells examined. $*P = 3.5 \times 10^{-5}$ (**s**), $P = 0.7907$ (NS) (**t**). **u**, Total ATP level in HeLa cells treated as indicated. Data are mean ± s.d. from three independent experiments. *P* values determined by two-tailed *t*-test.

Extended Data Table 1 | Proteins that interact with ARFGAP1

Abbreviation	Protein Name	MWT(kDa)
Tln1	Protein Tln1	270
Copa	Coatomer subunit alpha	138
Ap2b1	AP-2 complex subunit beta	105
Ap2a2	AP-2 complex subunit alpha-2	104
Copb2	Coatomer subunit beta'	102
Copg1	Coatomer subunit gamma-1	98
Arcn1	Coatomer subunit delta	57
Aldh2	Aldehyde dehydrogenase, mitochondrial	56
Fh	Fumarate hydratase, mitochondrial	54
Ctbp1	C-terminal-binding protein 1	47
Cth	Cystathionine gamma-lyase	44
Akr7a2	Aflatoxin B1 aldehyde reductase member 2	41
Taldo1	Transaldolase	37
Gapdh	Glyceraldehyde-3-phosphate dehydrogenase	36
Apoe	Apolipoprotein E	36
Mdh2	Malate dehydrogenase	36
Grhpr	Grhpr protein	36
Eif2s1	Eukaryotic translation initiation factor 2 subunit 1	36
Hadh	Hydroxyacyl-coenzyme A dehydrogenase, mitochondrial	34
Haao	3-hydroxyanthranilate 3,4-dioxygenase	33
Gnmt	Glycine N-methyltransferase	33
Qprt	Nicotinate-nucleotide pyrophosphorylase	31
Psma3	Proteasome subunit alpha type-3	28
Psme2	Proteasome activator complex subunit 2	27

Mass spectrometry was used to identify interacting proteins following a pull-down experiment using a GST fusion of ARFGAP1 and cytosol.

LETTER

https://doi.org/10.1038/s41586-018-0462-y

Cotranslational assembly of protein complexes in eukaryotes revealed by ribosome profiling

Ayala Shiber[1,2]*, Kristina Döring[1,2], Ulrike Friedrich[1,2], Kevin Klann[1,2], Dorina Merker[1,2], Mostafa Zedan[1,2], Frank Tippmann[1,2], Günter Kramer[1,2]* & Bernd Bukau[1,2]*

The folding of newly synthesized proteins to the native state is a major challenge within the crowded cellular environment, as nonproductive interactions can lead to misfolding, aggregation and degradation[1]. Cells cope with this challenge by coupling synthesis with polypeptide folding and by using molecular chaperones to safeguard folding cotranslationally[2]. However, although most of the cellular proteome forms oligomeric assemblies[3], little is known about the final step of folding: the assembly of polypeptides into complexes. In prokaryotes, a proof-of-concept study showed that the assembly of heterodimeric luciferase is an organized cotranslational process that is facilitated by spatially confined translation of the subunits encoded on a polycistronic mRNA[4]. In eukaryotes, however, fundamental differences—such as the rarity of polycistronic mRNAs and different chaperone constellations—raise the question of whether assembly is also coordinated with translation. Here we provide a systematic and mechanistic analysis of the assembly of protein complexes in eukaryotes using ribosome profiling. We determined the in vivo interactions of the nascent subunits from twelve hetero-oligomeric protein complexes of *Saccharomyces cerevisiae* at near-residue resolution. We find nine complexes assemble cotranslationally; the three complexes that do not show cotranslational interactions are regulated by dedicated assembly chaperones[5–7]. Cotranslational assembly often occurs uni-directionally, with one fully synthesized subunit engaging its nascent partner subunit, thereby counteracting its propensity for aggregation. The onset of cotranslational subunit association coincides directly with the full exposure of the nascent interaction domain at the ribosomal tunnel exit. The action of the ribosome-associated Hsp70 chaperone Ssb[8] is coordinated with assembly. Ssb transiently engages partially synthesized interaction domains and then dissociates before the onset of partner subunit association, presumably to prevent premature assembly interactions. Our study shows that cotranslational subunit association is a prevalent mechanism for the assembly of hetero-oligomers in yeast and indicates that translation, folding and the assembly of protein complexes are integrated processes in eukaryotes.

To test whether protein assembly in eukaryotes initiates during translation, we analysed 12 hetero-oligomeric complexes of *S. cerevisiae* (Extended Data Table 1). They were chosen to represent a variety of cellular functions, structural architectures, regulatory features, abundance and interface size. They are all verified complexes[3], mainly stable ones[3], with surface-exposed C termini for affinity tagging, and cytoplasmic or nuclear localization.

To identify the nascent-chain interaction profiles of complex subunits in vivo, we used selective ribosome profiling (SeRP)[9]. SeRP[9,10] compares the distribution of ribosome-protected mRNA footprints of two distinct samples generated from a single culture. One comprises the ribosome protected footprints of all translated open reading frames (ORFs) (total translatome). The other contains footprints of a selected set of ribosomes, co-purified with a tagged interaction partner (selected translatome). Accumulation of footprints in the selected translatome, as compared to the total translatome, directly indicates when it is during translation that the nascent chain interacts with the affinity-purified tagged protein subunit, at near-residue resolution.

We first analysed the assembly of fatty acid synthase (FAS), a multifunctional enzyme integrating all the fatty acid biosynthesis steps[11]. FAS is composed of two multi-domain subunits, α and β, which assemble to a highly intertwined, 2.6-MDa, hetero-dodecameric ($\alpha_6\beta_6$) complex[11] (Fig. 1a, d). To capture cotranslational assembly in vivo, we generated two strains, each chromosomally encoding one of the FAS subunits C-terminally fused to GFP for immunopurification. Tagging did not affect function (Extended Data Fig. 1a). SeRP demonstrates that FAS assembly initiates cotranslationally in a specific, asymmetric manner. Tagged α does not engage ribosome–nascent-chain complexes (RNCs) translating α or β. By contrast, tagged β engages RNCs synthesizing nascent α, leading to a strong, approximately 40-fold enrichment of selected footprints over total ribosome-protected footprints, starting near residue 125 of α and persisting until synthesis ends (Fig. 1b). This asymmetry of cotranslational interactions contrasts with immunoblotting results for the mature FAS, showing that each FAS subunit can immunopurify their partner subunit post-translationally with the same 1:1 stoichiometry (Extended Data Fig. 1b). The FAS subunits hence have distinct roles in the cotranslational assembly of the complex.

The onset of cotranslational subunit engagement directly correlates with FAS structural features: it coincides with ribosome exposure of the first 94 amino acids of α—which are intertwined with the last 389 amino acids of β—to form a single catalytic domain, the malonyl/palmitoyl-transferase (MPT) domain[11] (Fig. 1d). This implies that cotranslational assembly initiates upon formation of the MPT domain, the most stable interface between the two subunits[12]. To test whether the MPT interface is indeed required for cotranslational assembly of FAS, we analysed cotranslational interactions of FAS-deletion mutants lacking the MPT segments. Supporting the proposed model, MPT segments deletion, in either α or β, strongly reduces cotranslational interactions (Fig. 1c).

We tested whether cotranslational interactions are nascent-chain dependent by puromycin treatment, triggering the release of nascent chains from ribosomes[13]. Quantitative reverse transcription PCR (RT–qPCR) after immunopurifaction of the β subunit revealed that puromycin reduces the level of co-purified α-encoding mRNAs (Extended Data Fig. 1c, d), suggesting that cotranslational assembly relies on subunit association with nascent chains during translation. We next tested the extent of post-lysis association of β with nascent α and found it to be very low (Extended Data Fig. 1e–g). We conclude our SeRP setup provides snapshots of physiological interactions with RNCs that were established in vivo. Taken together, our findings indicate that the assembly of the dodecameric ($\alpha_6\beta_6$) FAS initiates cotranslationally by the formation of $\alpha\beta$ hetero-dimers, mediated by the interaction of the C terminus of β with the N terminus of nascent α to form the MPT domain (Extended Data Fig. 1h).

---

[1]Center for Molecular Biology of Heidelberg University (ZMBH), DKFZ-ZMBH Alliance, Heidelberg, Germany. [2]German Cancer Research Center (DKFZ), Heidelberg, Germany. *e-mail: a.shiber@zmbh.uni-heidelberg.de; g.kramer@zmbh.uni-heidelberg.de; bukau@zmbh.uni-heidelberg.de

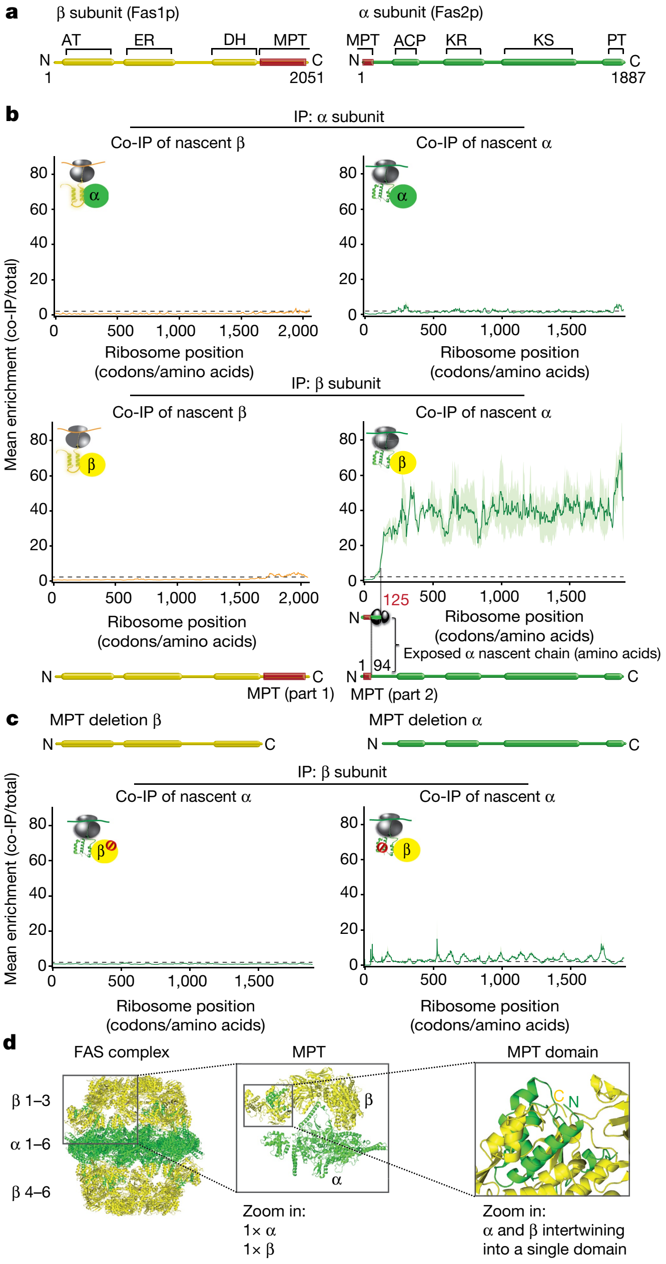


Fig. 1 | Cotranslational assembly of the FAS complex. a, Domain organization of FAS subunits: acyltransferase (AT), enoyl-reductase (ER), dehydratase (DH), malonyl/palmitoyl-transferase (MPT), acyl carrier protein (ACP), ketoreductase (KR), ketoacyl synthase (KS), phosphopantetheine transferase (PT). **b**, Nascent β and α engagement by C-terminally tagged α (top) or by C-terminally tagged β (bottom), analysed by SeRP. The ribosome position at which the enrichment stably crosses the twofold threshold (codon 125) is indicated. The area between replicates is shaded, indicating the degree of experimental variation. Data are from two biologically independent experiments. IP, immunopurification. **c**, Effect of the deletion of the MPT domain segment on cotranslational interactions, analysed as in **b**. Data are from two biologically independent experiments. **d**, Structural characteristics of the FAS complex and the MPT domain (PDB: 2UV8).

Our SeRP data correlate with the differential aggregation propensities of the individual FAS subunits. Upon exposure to various stresses, α becomes highly prone to aggregation and degradation, while β remains soluble[14,15]. Similarly, β remains stable in mutants lacking α, whereas α is rapidly degraded in mutants lacking β[16,17]. These findings support a model in which the structurally robust β folds independently, then serves as a scaffold to chaperone the cotranslational folding and assembly of the unstable α, protecting it from aggregation. Thus, cotranslational assembly may ameliorate the challenging folding trajectory of α.

We next analysed the assembly of a hetero-trimeric complex, the multi-aminoacyl-tRNA synthetase. This complex is composed of the essential methionyl- and glutamyl-tRNA synthetases MetRS and GluRS (encoded by *MES1* and *GUS1*, respectively), both of which are required for charging their specific tRNA with cognate amino acids, and the Arc1p cofactor, which regulates their catalytic activities and subcellular distributions[18–20] (Fig. 2a, d). We generated three strains, each chromosomally encoding one of the complex subunits C-terminally fused to GFP. Tagging did not affect function (Extended Data Fig. 2a). SeRP revealed that both GluRS and MetRS engage each other cotranslationally, resulting in at least a 30-fold enrichment in footprints, starting at codons 196 and 168 of *GUS1* and *MES1*, respectively, and persisting until synthesis ends. Both catalytic subunits also engage the nascent Arc1p cofactor, with nearly identical onsets approximately at codon 160 of *ARC1* (Fig. 2b). For all these nascent chains, the onset of partner subunit engagement occurs upon emergence at the ribosome exit tunnel of the N-terminal interaction domains, which share a similar glutathione-*S*-transferase (GST)-like fold[20]. Either catalytic subunit can thus cotranslationally engage all other subunits. By contrast, the fully synthesized Arc1p associates mainly with nascent GluRS (starting at codon 143) in a fluctuating manner, suggesting that these interactions are less stable compared to the catalytic subunits (Fig. 2b, lower panels). Our combined findings suggest that the assembly of multi-aminoacyl-tRNA synthetase initiates by cotranslational interactions of each of its subunits in a network-like manner (Extended Data Fig. 2b), which involves the shared GST-like folds acting as assembly drivers.

Notably, both GluRS and MetRS are bi-functional proteins regulating ATP-synthase expression upon glucose depletion. Arc1p is then rapidly degraded; MetRS relocates to the nucleus and GluRS relocates to the mitochondria[21]. As the localization signal of each of the two subunits is buried within the interface domains upon trimerization[21], we speculate that cotranslational assembly can regulate dual protein targeting in eukaryotes, by prioritizing cytosolic activity under favourable growth conditions.

To investigate the prevalence of the cotranslational assembly mechanism, we subjected 10 additional complexes to SeRP analysis. In total, 12 complexes composed of 26 individual subunits were analysed. We find that 9 out of 12 complexes exhibit cotranslational subunit interactions, demonstrating the prevalence of this assembly mechanism among stable cytosolic complexes (see PFK, TRP further examples in Extended Data Figs. 3, 4; Extended Data Table 2). Six out of nine complexes use a directional assembly mode, with one specific subunit being released from the ribosome before engaging the nascent interaction partner or partners (FAS, NatA, NatB, TRP, CPA and eIF2; Extended Data Table 2).

We hypothesized the cotranslationally engaged subunits have a higher propensity to misfold compared to their fully-synthesized partners. Accordingly, FAS subunits display asymmetric misfolding propensities[14–17]. To test whether this is a general feature, we performed in vivo aggregation and stability assays of subunits in wild-type and single subunit deletion strains for NatA, TRP and CPA. We excluded all complexes that are essential (eIF2)[22] or show severe growth phenotype upon subunit deletion (NatB)[23]. All nascently engaged subunits tested are indeed prone to aggregation or degradation in the absence of their partner subunits. By contrast, subunits that are only engaged after release from the ribosome are much more soluble and stable in the absence of their partner subunits (Extended Data Fig. 5a–c). Our findings suggest that in particular aggregation-prone subunits engage their partner subunits cotranslationally.

Three complexes do not show cotranslational assembly: (i) 20S proteasome subunits α1, 2; (ii) V-type-ATPase catalytic hexamer (A3, B3) and (iii) ribonucleotide reductase RNR (Rnr2p and Rnr4p complex). All three complexes are tightly controlled by dedicated

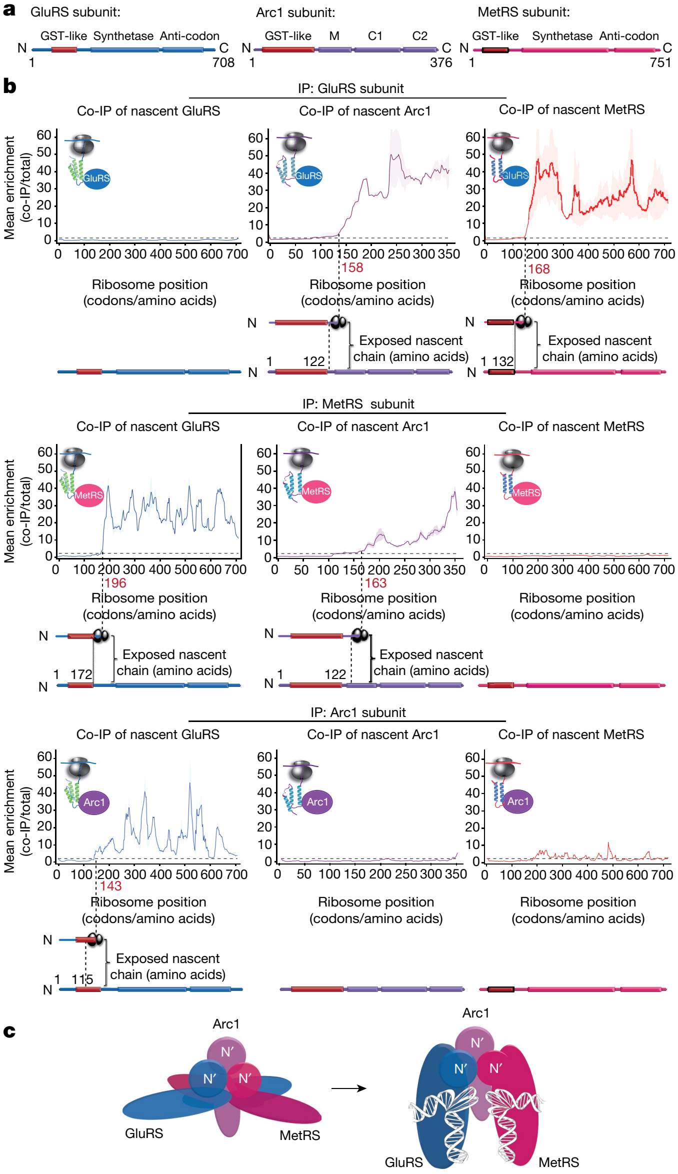


Fig. 2 | Cotranslational assembly of the aminoacyl-tRNA-synthetase complex. a, Domain organization of aminoacyl-tRNA-synthetase subunits. **b**, Engagement of nascent GluRS (left), Arc1 (middle) and MetRS (right) by C-terminally tagged GluRS (top), C-terminally tagged MetRS (middle) or C-terminally tagged Arc1 (bottom), analysed by SeRP. Coloured numbers indicate ribosome positions at which the enrichments stably cross the twofold threshold (dotted lines). The area between replicates area shaded, indicating the degree of experimental variation. Data are from two biologically independent experiments. **c**, Illustration of the subunits N′-terminal interfaces and structural fluctuations upon tRNA binding, based on structural data derived from a previous study[19].

assembly chaperones or inhibitors[5–7]. We speculate that these dedicated assembly factors function cotranslationally, protecting subunits from misfolding and premature binding to their partner subunits.

The position-resolved cotranslational interaction profiles of all 14 subunits identified in this study enabled us to reveal general features of the assembly process. We find that the onsets of interactions vary, but they are generally stable, persisting until synthesis ends (Fig. 3a, Extended Data Fig. 5d). Analysis of the nascent-chain features revealed that subunits containing extreme C-terminal interaction domains are excluded. In nearly all complexes, subunits are engaged when a complete interaction domain and an additional 24–37 amino acids have been synthesized (Fig. 3b). The eukaryotic ribosomal tunnel accommodates approximately 24 amino acids in extended conformation and approximately 38 amino acids in α-helical conformation[24]. Thus, the sharp onset of assembly (Fig. 3c) directly correlates with the emergence of the entire interface domain from the ribosome exit tunnel. Taken together, our results suggest assembly is facilitated by interface domains cotranslational folding.

Folding of nascent polypeptides in yeast is facilitated by the Hsp70 family member Ssb, the major ribosome-associated chaperone[8,10,25]. Ssb is targeted to the ribosome by the RAC complex[25] and by direct contacts with the exit tunnel[26], ensuring high affinity to short, hydrophobic nascent-chain segments[10]. This raises the question of how Ssb binding relates to cotranslational complex assembly. Analysis of Ssb SeRP interaction profiles[10] shows that all nascent chains that engage partner subunits have one or multiple Ssb binding peaks. Ssb binds 13 out of 14 subunits before the onset of cotranslational assembly, generally during the synthesis of interaction domains, and dissociates just before subunit engagement (Fig. 4b, c for examples; Fig. 4d, e). Ssb engagement is thus well coordinated with assembly. We propose that Ssb shields hydrophobic patches within interaction domains, protecting them from non-productive interactions and misfolding. Ssb dissociates upon full ribosome exposure of these domains, permitting cotranslational folding and subunit joining. We further investigated Ssb interplay with assembly by a proteome-wide bioinformatics analysis, identifying all putative cotranslationally assembled subunits (for details, see 'Data analysis section' in Methods). Metagene profiling of Ssb binding to these ORFs and/or nascent chains demonstrates that Ssb generally dissociates just before putative cotranslational assembly-onset positions, which are characterized by low hydrophobicity (Extended Data Fig. 6). We suggest that the low hydrophobicity disfavours Ssb binding, allowing for interface domain folding and subunit interaction (see conclusions for model). To directly assess the effect of Ssb on cotranslational assembly, we attempted SeRP experiments in *ssb1Δssb2Δ* cells. However, these experiments repeatedly failed owing to the low amounts of ribosomes co-purified with tagged subunits. Nevertheless, these results are consistent with Ssb having an important role in cotranslational assembly. Accordingly, *ssb1Δssb2Δ* mutants display widespread aggregation of newly synthesized proteins, among which complex subunits are enriched—including most of the complex subunits analysed here[27] (Extended Data Table 2).

Beyond complex assembly, we hypothesized that cotranslational interactions may extend to all protein–protein networks. We tested this possibility by identifying the proteome-wide nascent-chain interactions of some subunits in our dataset, focusing on the subunits of enzymatic pathways. We adapted a recently developed peak detection algorithm[10], to identify local binding peaks, which were defined as a greater than threefold enrichment in footprint density over a stretch of more than ten codons. For FAS β, PFK β and Cpa2 subunits we detected additional, transient interactions with distinct sets of RNCs known to be functionally related to or directly interacting with the subunit (examples in Extended Data Fig. 7). One example is FAS β, which engages nascent acetyl-CoA carboxylase (Acc1p). Acc1p catalyses the step directly preceding FAS in the pathway (Extended Data Fig. 7a). Unlike the stable engagement of FAS β with nascent α for assembly, its association with nascent Acc1p is transient, similar to the interactions between fully synthesized FAS and Acc1p that have previously been reported[28]. Nonetheless, it is specific, as β does not engage any other nascent member of the fatty acid synthesis pathway (Extended Data Fig. 7a). These findings provide first evidence that metabolic pathways can be coordinated cotranslationally. The extent and function of such nascent-chain interactomes have yet to be revealed.

To conclude, our study provides direct in vivo evidence, at near-residue resolution, that cotranslational subunit engagement is a widespread mechanism for complex assembly in eukaryotes. Our findings are consistent with previous studies that used indirect approaches to study cotranslational interactions in eukaryotes, such as RNA-IP-microarray (RIP-Chip)[29,30] or in vitro translation system[31].

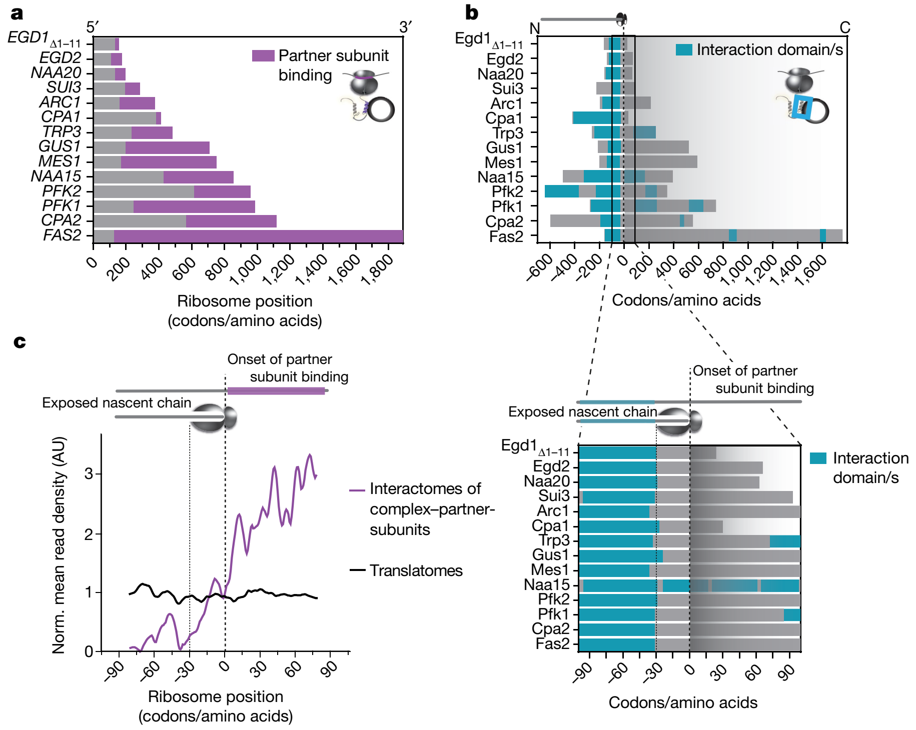


Fig. 3 | Characteristics of cotranslational complex assembly interactions. a, Onset and persistence of the cotranslational interaction of each subunit with its partner complex subunit or subunits, for all 14 subunits identified as cotranslationally engaged. *NAA20* and *NAA15* are also known as *NAT3* and *NAT1*, respectively. **b**, Top, interaction domain exposure correlated to the onset of assembly onset. Bottom, an expanded view of the region surrounding the onset of assembly. **c**, Normalized mean read density of interaction profiles of 14 cotranslationally engaged subunits, aligned and zoomed-in to the onset of assembly of each nascent chain. AU, arbitrary units.

The high misfolding propensities of the subunits that interact as nascent chains with partner subunits underscore the importance of this mechanism. Cotranslational assembly may be a prerequisite for the evolvement of complex folding architectures and the rescue subunits destabilized by accumulating mutations. We furthermore reveal an intricate functional interplay between the Ssb chaperone and

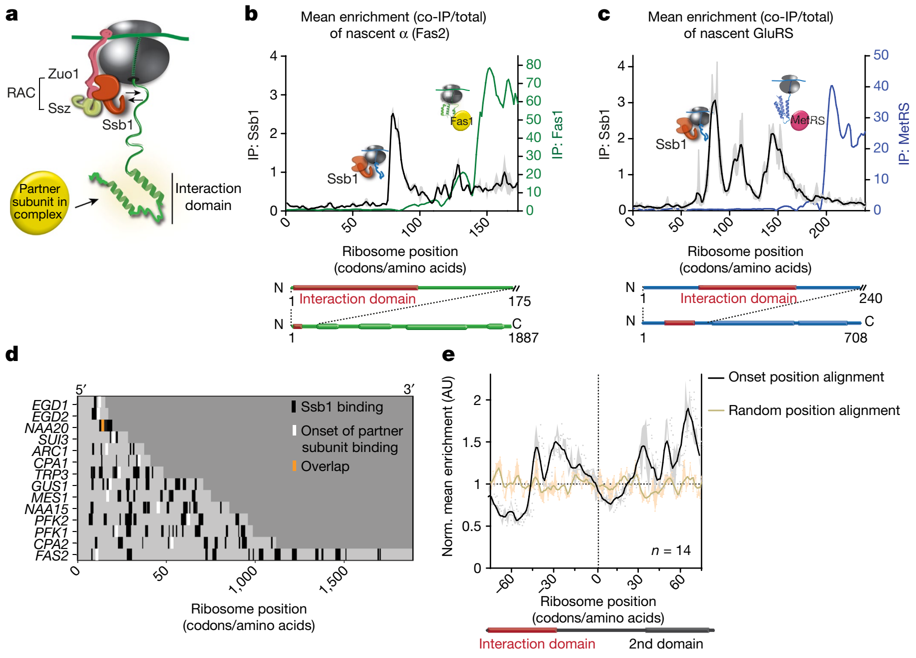


Fig. 4 | Coordination of cotranslational complex assembly with the ribosome-associated chaperone Ssb binding. a, Illustration of ribosome–nascent-chain binding to Ssb or a partner subunit. **b**, **c**, Zoomed-in interaction profiles of Ssb1–GFP and cotranslationally engaged partner subunits with the nascent FAS α (**b**, Fas2) and nascent GluRS (**c**), analysed by SeRP. The area between replicates is shaded, indicating the degree of experimental variation. **d**, Heat map of Ssb1–GFP binding to ribosomes synthesizing the 14 cotranslationally engaged subunits, compared to complex assembly onset. **e**, Metagene analysis of the Ssb1–GFP interaction profiles with 14 cotranslationally engaged nascent chains, aligned and zoomed-in to assembly onset, compared to random position along the ORFs alignment. There is no correlation between the onset and random position alignment (Pearson correlation $r = 0.01256$), thus Ssb depletion at onset positions is significant. The area between replicates is shaded, indicating the degree of experimental variation. **b**–**e**, Data are from two biologically independent experiments.

the binding of partner subunits, suggesting that nascent subunits are constantly engaged (for model, see Extended Data Fig. 8). Conversely, exposed interfaces may serve as signals for subunit degradation, providing a molecular basis for quality control and the regulation of subunit stoichiometry at the level of the nascent chain. We further speculate that the translation of complex subunits is spatially confined in the cytosol, as this would facilitate timely assembly and prevent prolonged nascent-chain exposure.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0462-y

Received: 11 November 2017; Accepted: 2 August 2018;
Published online 29 August 2018.

1. Balchin, D., Hayer-Hartl, M. & Hartl, F. U. In vivo aspects of protein folding and quality control. *Science* **353**, aac4354 (2016).
2. Gloge, F., Becker, A. H., Kramer, G. & Bukau, B. Co-translational mechanisms of protein maturation. *Curr. Opin. Struct. Biol.* **24**, 24–33 (2014).
3. Benschop, J. J. et al. A consensus of core protein complex compositions for *Saccharomyces cerevisiae*. *Mol. Cell* **38**, 916–928 (2010).
4. Shieh, Y. W. et al. Operon structure and cotranslational subunit association direct protein assembly in bacteria. *Science* **350**, 678–680 (2015).
5. Le Tallec, B. et al. 20S proteasome assembly is orchestrated by two distinct pairs of chaperones in yeast and in mammals. *Mol. Cell* **27**, 660–674 (2007).
6. Smardon, A. M., Tarsio, M. & Kane, P. M. The RAVE complex is essential for stable assembly of the yeast V-ATPase. *J. Biol. Chem.* **277**, 13831–13839 (2002).
7. Meurisse, J. et al. Hug1 is an intrinsically disordered protein that inhibits ribonucleotide reductase activity by directly binding Rnr2 subunit. *Nucleic Acids Res.* **42**, 13174–13185 (2014).
8. Pfund, C., Huang, P., Lopez-Hoyo, N. & Craig, E. A. Divergent functional properties of the ribosome-associated molecular chaperone Ssb compared with other Hsp70s. *Mol. Biol. Cell* **12**, 3773–3782 (2001).
9. Becker, A. H., Oh, E., Weissman, J. S., Kramer, G. & Bukau, B. Selective ribosome profiling as a tool for studying the interaction of chaperones and targeting factors with nascent polypeptide chains and ribosomes. *Nat. Protocols* **8**, 2212–2239 (2013).
10. Doring, K. et al. Profiling Ssb–nascent chain interactions reveals principles of Hsp70-assisted folding. *Cell* **170**, 298–311.e220 (2017).
11. Leibundgut, M., Jenni, S., Frick, C. & Ban, N. Structural basis for substrate delivery by acyl carrier protein in the yeast fatty acid synthase. *Science* **316**, 288–290 (2007).
12. Gipson, P. et al. Direct structural insight into the substrate-shuttling mechanism of yeast fatty acid synthase by electron cryomicroscopy. *Proc. Natl Acad. Sci. USA* **107**, 9164–9169 (2010).
13. Hansen, W. J., Lingappa, V. R. & Welch, W. J. Complex environment of nascent polypeptide chains. *J. Biol. Chem.* **269**, 26610–26613 (1994).
14. Cherkasov, V. et al. Systemic control of protein synthesis through sequestration of translation and ribosome biogenesis factors during severe heat stress. *FEBS Lett.* **589**, 3654–3664 (2015).
15. Duttler, S., Pechmann, S. & Frydman, J. Principles of cotranslational ubiquitination and quality control at the ribosome. *Mol. Cell* **50**, 379–393 (2013).
16. Schüller, H. J., Förtsch, B., Rautenstrauss, B., Wolf, D. H. & Schweizer, E. Differential proteolytic sensitivity of yeast fatty acid synthetase subunits α and β contributing to a balanced ratio of both fatty acid synthetase components. *Eur. J. Biochem.* **203**, 607–614 (1992).
17. Scazzari, M., Amm, I. & Wolf, D. H. Quality control of a cytoplasmic protein complex: chaperone motors and the ubiquitin-proteasome system govern the fate of orphan fatty acid synthase subunit Fas2 of yeast. *J. Biol. Chem.* **290**, 4677–4687 (2015).
18. Karanasios, E., Simader, H., Panayotou, G., Suck, D. & Simos, G. Molecular determinants of the yeast Arc1p–aminoacyl-tRNA synthetase complex assembly. *J. Mol. Biol.* **374**, 1077–1090 (2007).
19. Koehler, C., Round, A., Simader, H., Suck, D. & Svergun, D. Quaternary structure of the yeast Arc1p–aminoacyl-tRNA synthetase complex in solution and its compaction upon binding of tRNAs. *Nucleic Acids Res.* **41**, 667–676 (2013).
20. Simader, H. et al. Structural basis of yeast aminoacyl-tRNA synthetase complex formation revealed by crystal structures of two binary sub-complexes. *Nucleic Acids Res.* **34**, 3968–3979 (2006).
21. Frechin, M. et al. Expression of nuclear and mitochondrial genes encoding ATP synthase is synchronized by disassembly of a multisynthetase complex. *Mol. Cell* **56**, 763–776 (2014).
22. Hannig, E. M., Cigan, A. M., Freeman, B. A. & Kinzy, T. G. GCD11, a negative regulator of GCN4 expression, encodes the gamma subunit of eIF-2 in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **13**, 506–520 (1993).
23. Polevoda, B., Cardillo, T. S., Doyle, T. C., Bedi, G. S. & Sherman, F. Nat3p and Mdm20p are required for function of yeast NatB Nα-terminal acetyltransferase and of actin and tropomyosin. *J. Biol. Chem.* **278**, 30686–30697 (2003).
24. Bhushan, S. et al. α-Helical nascent polypeptide chains visualized within distinct regions of the ribosomal exit tunnel. *Nat. Struct. Mol. Biol.* **17**, 313–317 (2010).
25. Gautschi, M., Mun, A., Ross, S. & Rospert, S. A functional chaperone triad on the yeast ribosome. *Proc. Natl Acad. Sci. USA* **99**, 4209–4214 (2002).
26. Gumiero, A. et al. Interaction of the cotranslational Hsp70 Ssb with ribosomal proteins and rRNA depends on its lid domain. *Nat. Commun.* **7**, 13563 (2016).
27. Willmund, F. et al. The cotranslational function of ribosome-associated Hsp70 in eukaryotic protein homeostasis. *Cell* **152**, 196–209 (2013).
28. Gavin, A. C. et al. Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440**, 631–636 (2006).
29. Duncan, C. D. & Mata, J. Widespread cotranslational formation of protein complexes. *PLoS Genet.* **7**, e1002398 (2011).
30. Halbach, A. et al. Cotranslational assembly of the yeast SET1C histone methyltransferase complex. *EMBO J.* **28**, 2959–2970 (2009).
31. Sakahira, H. & Nagata, S. Co-translational folding of caspase-activated DNase with Hsp70, Hsp40, and inhibitor of caspase-activated DNase. *J. Biol. Chem.* **277**, 3364–3370 (2002).

Acknowledgements We thank members of the Bukau laboratory for discussions, D. H. Wolf for FAS antisera and the DKFZ Core facility for sequencing. This work was supported by ERC Advanced grant (743118), SFB1036 and an Alexander von Humboldt fellowship to A.S.

Reviewer information Nature thanks J. Marsh and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions A.S., G.K. and B.B. conceived the study and designed the experiments. A.S., K.D., U.F., K.K., D.M. and M.Z. performed the experiments. A.S., K.D., U.F., K.K., D.M., M.Z., F.T., G.K. and B.B. analysed the data. A.S., G.K. and B.B. wrote the manuscript with input from all authors.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0462-y.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0462-y.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to A.S. or G.K. or B.B.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Strains construction. GFP-tagged strains and deletion strains were generated via homologous recombination, constructed according to previously published work[32]. For the GFP tag, a cassette containing the monomeric GFP gene and a G418 resistance marker was amplified from the pYM12-mGFP plasmid. For gene deletions, a cassette containing only a selection marker was PCR amplified. All experiments were performed in the BY4741 strain background. *S. cerevisiae* strains used in this study are listed in Supplementary Table 1.

Yeast cultures. Yeast cultures were cultivated either in liquid yeast extract–peptone–dextrose (YPD)-rich medium, or in synthetic dextrose (SD) minimal medium (1.7 g/l yeast nitrogen base with ammonium sulfate or 1.7 g/l yeast nitrogen base without ammonium sulfate with 1 g/l monosodium glutamic acid, 2% glucose and supplemented with a complete or appropriate mixture of amino acids) at 30 °C. Trp2-GFP, Trp3-GFP strains were grown in SD lacking tryptophan; and Cpa1-GFP, Cpa2-GFP were grown in SD lacking arginine, to induce their expression. For fatty acid supplementation, SD media was supplemented with 0.03% Myristic acid (Sigma, pre-dissolved in DMSO), 0.1% Tween-40 (Sigma), and 0.05% yeast extract.

Purification of RNCs for SeRP. Approximately 800 ml of cell culture was grown to an $OD_{600\,nm}$ of 0.5, at 30 °C, in appropriate medium. Cell collection was performed in the culture medium as follows: cells were collected rapidly by vacuum filtration on 0.45-μm nitrocellulose (Aamersham) blotting membrane and then flash frozen, as previously described[10]. Next, cells were lysed by cryogenic grinding in a mixer mill (2 min, 30 Hz, MM400 Retsch) with 900 μl of lysis buffer (20 mM Tris–HCl pH 8.0, 140 mM KCl, 6 mM $MgCl_2$, 0.1% NP-40, 0.1 mg/ml cycloheximide (CHX), 1 mM PMSF, 2× protease inhibitors (Complete EDTA-free, Roche), 0.02 U/ml DNaseI (recombinant DNaseI, Roche), 20 mg/ml leupeptin, 20 mg/ml aprotinin, 10 mg/ml E-64, 40 mg/ml bestatin). Lysates were cleared by centrifugation (2 min at 30,000*g*, 4 °C).

For each experiment, supernatants were divided for total (200 μl) and immunopurification (700 μl) translatome samples. Total samples were digested using 10 U per $A_{260\,nm}$ unit of RNaseI for 25 min at 4 °C, in rotation, then loaded onto 800 μl of sucrose cushions (25% sucrose, 20 mM Tris-HCl pH 8.0, 140 mM KCl, 10 mM $MgCl_2$, 0.1 mg/ml CHX, 1× protease inhibitors) and centrifuged in a TLA120-rotor for 90 min at 75,000 rpm, 4 °C. Pellets were resuspended in lysis buffer and transferred to non-stick tubes. 100–200 mg of total RNA were taken for ribosome profiling of the total translatome.

Immunopurification samples were digested using 10 U per $A_{260\,nm}$ unit of RNaseI, together with 100–400 μl of GFP-binder slurry and the suspension was rotated for 25 min, 4 °C. Beads were washed three times in wash buffer I (20 mM Tris-HCl pH 8.0, 140 mM KCl, 10 mM $MgCl_2$, 1 mM PMSF, 0.1% NP-40, 0.1 mg/ml CHX, 2× protease inhibitors) (3 min, 3 × 1 min) and twice in wash buffer II (20 mM Tris-HCl, 140 mM KCl, 10 mM $MgCl_2$, 1 mM PMSF, 0.1 mg/ml CHX, 0.01% NP-40, 10% glycerol, 2× protease inhibitors) (5 min, once 1 min and again for 4 min). The washed beads were subsequently used for RNA or protein extraction. Affinity purification was analysed by western blot with aliquots of each step.

cDNA library preparation for deep sequencing. Library preparation was performed mainly as described[10]. In summary, RNA extraction was performed by mixing 0.75 ml pre-warmed acid phenol (Ambion) with either the purified monosomes of the total translatome or the monosomes bound to affinity beads for the immunopurification translatomes and 40 ml 20% SDS (Ambion). After shaking at 1,400 rpm for 5 min at 65 °C, samples were incubated 5 min on ice and centrifuged at 20,000*g* for 2 min. Top aqueous layers were transferred to fresh tubes and mixed again with 0.7 ml acid phenol. Samples were incubated for 5 min at room temperature with occasional vortexing and afterward centrifuged for 2 min at 20,000*g*. Top aqueous layers were transferred to fresh tubes and mixed with 0.6 ml chloroform, vortexed and centrifuged for 1 min at 20,000*g*. Nucleic acids were precipitated by adding 78 ml 3 M NaOAc pH 5.5, 2 ml glycoblue and 0.75 ml isopropanol and incubating for 1 h to 16 h at −20 °C. Samples were centrifuged for 30 min at 20,000*g*, 4 °C and pellets were washed with ice-cold 80% ethanol and resuspended in 10 mM Tris–HCl pH 7.0. Samples were heated at 80 °C for 2 min and for total translatome 50 mg of RNA and for the immunopurification translatome the entire sample was loaded onto a 15% TBE–Urea polyacrylamide gel (Invitrogen) in 1× TBE (Ambion) and run for 65 min at 200 V. Gels were stained for 20 min with SYBR gold (Invitrogen). To recover ribosomal footprints, the gel pieces that contained RNA fragments were excised, with a size between 25 and 33 nucleotides. Gel pieces were placed into 0.5-ml gel breaker tubes, nested into a 1.5 ml tube and centrifuged for 3 min at 20,000*g*. Ten millimolar Tris-HCl pH 7.0 (0.5 ml) was added and tubes were incubated at 70 °C for 10 min with maximal shaking in an Eppendorf thermomixer. Gel pieces were removed using a Spin-X cellulose acetate column (Fisher) and the flow through was transferred to a new tube. Fifty-five millilitres of 3 M NaOAc pH 5.5, 2 ml glycoblue and 0.55 ml isopropanol were added. After mixing, tubes were frozen at −20 °C for 16 h. Samples were centrifuged for 30 min at 20,000*g* and 4 °C and pellets were washed with ice-cold 80% ethanol and resuspended in 15 ml of 10 mM Tris-HCl pH 7.0. For dephosphorylation, 2 μl 10× T4 polynucleotide kinase buffer without ATP (NEB), 1 ml murine RNase inhibitor and 2 μl T4 polynucleotide kinase (NEB) were added to each sample. Samples were incubated at 37 °C for 1 h before heat inactivation of the enzyme for 10 min at 75 °C precipitation of nucleic acids by adding 0.5 ml 10 mM Tris-HCl pH 7.0, 55 μl 3 M NaOAc pH 5.5, 2 μl glycoblue and 0.55 μl isopropanol and incubating for 1 h to 16 h at −20 °C. Samples were centrifuged for 30 min at 20,000*g* and 4 °C, pellets were washed with ice-cold 80% ethanol and resuspended in 6–11 μl of 10 mM Tris-HCl pH 7.0. For linker ligation, a maximum of 5 pmol RNA in 5 μl were denatured for 2 min at 80 °C before 8 μl 50% sterile filtered PEG MW 8000, 2 μl DMSO, 2 μl 10× T4 RNA Ligase 2 buffer (NEB), 1 μl murine RNase inhibitor, 1 μl 1 mg/μl linker L1 and 1 μl truncated T4 RNA Ligase 2 (NEB) were added and incubated for 2.5 h at 37 °C or 23 °C. Nucleic acids were precipitated as described before and resuspended in 6 μl 10 mM Tris-HCl pH 7.0. Samples were run on a 10% TBE–Urea polyacrylamide gel (Invitrogen) in 1× TBE (Ambion) for 50 min at 200 V. Gels were stained for 20 min with SYBR gold and desired gel pieces were excised and RNA was extracted as described before. For reverse transcription, RNA was resuspended in 10 μl 10 mM Tris-HCl pH 7.0 and 1 μl 10 mM dNTP (NEB), 1 μl 25 linker L1′L20 and 1.5 μl DEPC H20 were added to each sample. Samples were incubated at 65 °C for 5 min followed by addition of 4 μl 5× FSB buffer (Invitrogen), 1 μl murine RNase inhibitor, 1 μl 0.1 M DTT (Invitrogen) and 1 μl Superscript III (Invitrogen). Samples were incubated at 50 °C for 30 min and afterward 2.3 μl 1 N NaOH was added to hydrolyze RNA and samples were further incubated at 95 °C for 15 min. Samples were run on a 10% TBE–Urea polyacrylamide gel for 70 min at 200 V. Gels were stained as described before and desired bands were excised and nucleic acids were extracted as mentioned earlier but using Tris-HCl pH 8.0 and precipitating nucleic acids by adding 32 μl 5 M NaCl, 1 μl 0.5 M EDTA, 2 μl glycoblue and 0.55 μl isopropanol. For circularization, DNA was resuspended in 15 μl 10 mM Tris-HCl pH 8.0 and 2 μl 10× CircLigase buffer (EPICENTRE), 1 μl 1 mM ATP, 1 μl 50 mM $MnCl_2$ and 1 μl CircLigase (EPICENTRE) were added. Samples were incubated at 60 °C for 1 h. Addition of 1 μl CircLigase was repeated and samples were incubated for another hour at 60 °C. Afterward, the enzyme was inactivated by incubating 10 min at 80 °C. Five microlitres of circularized DNA was used for PCR amplification. Therefore, 16.7 μl 5× HF buffer, 1.7 μl 10 mM dNTPs, 0.4 μl 100 mMP CR primer L1′, 0.4 μl 100 mM barcoding primer, 59.2 μl DEPC H20 and 0.8 μl HF Phusion (NEB) were added. 17 μl PCR mix were aliquoted to four separate PCR tubes and the following PCR reaction cycles were run: 1.) 98 °C, 30 s; 2.) 98 °C, 10 s; 3.) 60 °C, 10 s; 4.) 72 °C, 5 s. Steps 2 through 4 were repeated ten times and one tube was removed after cycles 7–13. Samples were run on 8% TBE polyacrylamide gels (Invitrogen) in 1× TBE (Ambion) for 45 min at 180 V. Gels were stained as mentioned before and desired bands were excised and DNA was extracted as described before. After a quality control step using a high sensitivity Bioanalyzer chip (Agilent), samples were sequenced on a HiSeq 2000 (Illumina).

Data analysis. Sequenced reads were processed as previously described[10] using standard trimming and genome alignment tools (Cutadapt, Bowtie2, Tophat2) and Python scripts adapted to *S. cerevisiae*. SeRP analyses are based on at least two independent biological replicates that were highly reproducible, as evaluated by Pearson correlation analysis of each gene profile, see details below.

Ratio-based enrichment profiles analysis. The ratio-based enrichment profiles were built by comparing the RPM (reads per million mapped reads) interactome and translatome data at each nucleotide along the ORFs. The reproducibility of replicates of interaction profiles was evaluated by Pearson correlation analysis. If a threshold of 0.6 was passed, genes were processed further.

To exclude genes that are expressed close to the background level or have a low read coverage, we defined minimal requirement thresholds that must be all passed before genes were considered for analysis: (i) at least 64 reads in both subunit-bound and total translatome datasets; (ii) at least 8 RPKM (reads per kilobase of transcript, per million mapped reads) in both translatome datasets; (iii) at least one position after the first 90 nucleotides in the subunit-bound translatome that has a twofold higher read number than the average of the first 90 nucleotides (designated 90 nucleotides background giving the specific background signal for every gene; for genes lacking any read in the 90 nucleotide background, the average read per nucleotide along the complete gene from the corresponding translatome is used).

Enrichment threshold: local footprint density over a sequence stretch of minimal 90 nucleotides must be at least twofold enriched at every position.

Metagene analysis. For metagene analyses genes were normalized to their expression level by dividing the read density of each nucleotide by the average read density per nucleotide of the respective gene. Replicates reproducibility of gene profiles was evaluated by Pearson correlation analysis. If a threshold of 0.6 was passed, genes were processed further.

The genes were aligned to the position of onset of cotranslational complex assembly interactions, defined by the single codon position where the enrichment threshold (defined in the previous segment) was crossed for each gene.

Proteome-wide bioinformatics analysis of Ssb1 interplay with putative cotranslational complex assembly interactions:

To determine the proteome-wide interplay of Ssb SeRP interaction profiles[2] with cotranslational assembly interactions, we performed a bioinformatics analysis to identify all hetero-oligomeric complexes subunits that are putatively cotranslationally assembling, and their putative assembly onset positions.

For the analysis we have used the following parameters, extrapolated from our experimental data: of all PDB-deposited structures of hetero-oligomeric complexes, we identified subunits containing N′ terminal protein interface domains, located in the first 40% of genes, as our experimental dataset is disenriched for subunits containing extreme C′-terminal ones. Complexes involving interface domains smaller than five interacting residues were removed, as our experimental dataset is disenriched for this type of complex. Complexes involving dedicated assembly chaperones/inhibitors were excluded, as our experimental dataset shows this type of complex is less likely to cotranslationally assemble.

We next identified ends of N′ terminal interfaces as assembly onset positions, as our experimental data shows most assembly onsets occur directly upon the emergence of an entire interface domain from the ribosome exit tunnel.

We used a geometric clustering algorithm (http://www.blopig.com/blog/2013/10/get-pdb-intermolecular-protein-contacts-and-interface-residues/ from the Oxford Protein Informatics Group) to identify the patches of interface atoms within each subunit. Atoms directly involved in protein–protein interfaces were defined by an intermolecular distance cutoff of 4.5 Å, on the basis of previously published works[33–35].

Interface patch atoms were defined by an intramolecular distance cutoff of 10 Å, according to their C_α distances within the crystal component, on the basis of previously published work[33,34].

Interface patches size threshold. Patches must include at least five interacting residues. Structures with resolution greater than 10 Å were removed from the analysis, in accordance with these defined interface parameters, as the cutoff of 10 Å could not be determined. Membrane protein complexes were removed from the analysis. Structures of truncated proteins lacking their N′ terminal interfaces were removed from our analysis.

We aligned all proteins to the position of putative onset of cotranslational complex assembly interactions: to the ends of N′ terminal interfaces and performed a metagene profile of Ssb1 binding to this subset with Ssb1 SeRP experiments[10].

ORFs with low Ssb1 SeRP footprint coverage were removed from the analysis, using a threshold of 64 total counts per ORF to maintain significant reproducibility between SeRP independent biological replicates[10,36]. The reproducibility of replicates of interaction profiles was evaluated by Pearson correlation analysis. If a threshold of 0.6 was passed, genes were processed further.

In total, 116 subunits were identified as putatively cotranslationally assembling.

Software and algorithms. Bowtie2 v.2.2.5.0 (http://bowtie-bio.sourceforge.net/bowtie2/index.shtml). Tophat2 v.2.0.13 (http://ccb.jhu.edu/software/tophat/downloads/). Python v.2.7 and v.3.4 (https://www.python.org/downloads/).

Immunoblotting. Samples were dissolved in standard sample buffer and boiled at 95 °C for 5 min. Samples were separated on SDS–PAGE gels (4–12% gradient), transferred to polyvinylidene fluoride membranes and immunoblotted. The following antibodies were used: polyclonal rabbit FAS antibody[16] (a gift from D. H. Wolf), polyclonal rabbit GFP antibody (antiserum from rabbit raised against YFP)[37]. Proteins were visualized by enhanced chemi-fluorescence reaction.

Imaging. Cells were grown in SC medium containing 2% glucose, 30 °C to log phase. For anthranilate synthase subunits co-staining, cells were transferred to growth in SD lacking tryptophan for the last 30 min to induce their expression; similarly for carbamoyl phosphate synthetase subunits co-staining, cells were transferred to growth in SD lacking arginine, for the last 30 min to induce their expression.

Cells were fixed with 37% formaldehyde for 15 min, centrifuged at 1,200*g* for 12 min and resuspended in 4% paraformaldehyde and 100 mM KPO_4 at room temperature for 1 h.

High-sensitivity confocal imaging was performed on a Leica DMi8 spinning disk system with a Yokogawa CSU-X1 scanner unit. Images were acquired by using a HC PL APO 63×/1.40–0.60 oil objective lens (Leica), the Orca Flash 4.0 LT sCMOS camera (Hamamatsu, C11440-42U) and a quad band filter set and up to four diode laser lines (405 nm, 488 nm, 561 nm, 635 nm) with the MetaMorph Advanced Acquisition software (v.7.8.13.0, by Molecular Devices, was used for confocal imaging). *Z* stacks (0.2-μm steps) images were acquired for the 488-nm channel. All further processing of acquired images was performed with ImageJ software. A maximal projection of ~3–5 *Z* stacks is shown. For the purpose of foci quantification of subunits fused to GFP, both manual and automated ('FindFoci' open-source plugin for ImageJ[38]) quantifications were performed. Approximately 150 cells were analysed per sample with a total of three repetitions.

Quantitative PCR of mRNAs co-purified with GFP-tagged proteins. *GFP affinity purification.* GFP affinity purification was performed as described in above (see 'Purification of RNCs for SeRP'), for immunoprecipitation samples, with the following changes: no RNaseI treatment was performed and the lysis buffer was supplemented with KCl to a final concentration of 500 mM. For puromycin treatment samples, the lysis buffer was first supplemented with puromycin (10 μg ml; Invitrogen) and then added to the filtered cells. All subsequent lysis and immunoprecipitation steps were performed in the presence of puromycin. Samples were then directly subjected to phenol RNA extraction, as described[10].

Reverse transcription. First strand cDNA synthesis for quantitative PCR (qPCR) was performed using the Superscript III First Strand RT PCR kit (Invitrogen). One microgram of isolated RNA was mixed with 5 ng/μl random hexameric primers, 1 mM dNTPs, adjusted to 10 μl and incubated at 65 °C for 10 min and then chilled on ice. To the RNA–primer mix a premixed cDNA synthesis mix was added (2 μl 10× reverse transcription buffer, 4 μl 25 mM $MgCl_2$, 2 μl 100 mM DTT, 20 U RNaseOUT, 100 U Superscript III). Reaction was incubated for 50 min at 50 °C in a water bath and terminated by heating the mix to 85 °C for 5 min. After cooling on ice 0.5 μl RNase H were added and incubated at 37 °C for 20 min. cDNA then was stored at −80 °C or used directly for qPCR.

qPCR. qPCR was performed using the DyNAmo Flash SYBR Green qPCR Kit (Thermo Scientific) and a LightCycler 480 (Roche).

Reactions were pipetted in 384-well LightCycler480 multiwell plates (Roche). Per reaction 2.5 μl of cDNA (in appropriate dilution) was mixed with 7.5 μl reaction Master Mix (5 μl Flash SYBR Green Mix, 1.7 μl DEPC H_2O, 0.4 μl per primer (10 mM)) with a multistep pipette to reduce pipetting errors. For analysis the following program was used: pre-incubation: 95 °C, 5 min; amplification: 95 °C, 10 s, 55 °C, 20 s, 72 °C, 20 s; single acquisition mode; melting curve: 60–90 °C, 0.11 °C/s, continuous acquisition mode. C_p values were calculated by derivation by the LightCycler480 software (Roche). For normalization *ACT1* mRNA was used as a housekeeping gene.

CHX chase and flow cytometry analysis. Yeast cells were grown to log phase, then CHX (0.5 mg/ml) was added, and aliquots from each time point were taken. GFP levels of fixed cells at each time point were determined by flow cytometry analysis performed using a BD FACS Canto II equipped with lasers (405 nm, 488 nm, 635nm). Detectors used: FSC, SSC, 488-E for GFP with filter 530/30. Cell population gated on FSC/SSC area dot plot, exclusion of debris and cell aggregates by SSC/FSC height and width. Twenty thousand events(cells)/sample. Data are from three biologically independent experiments.

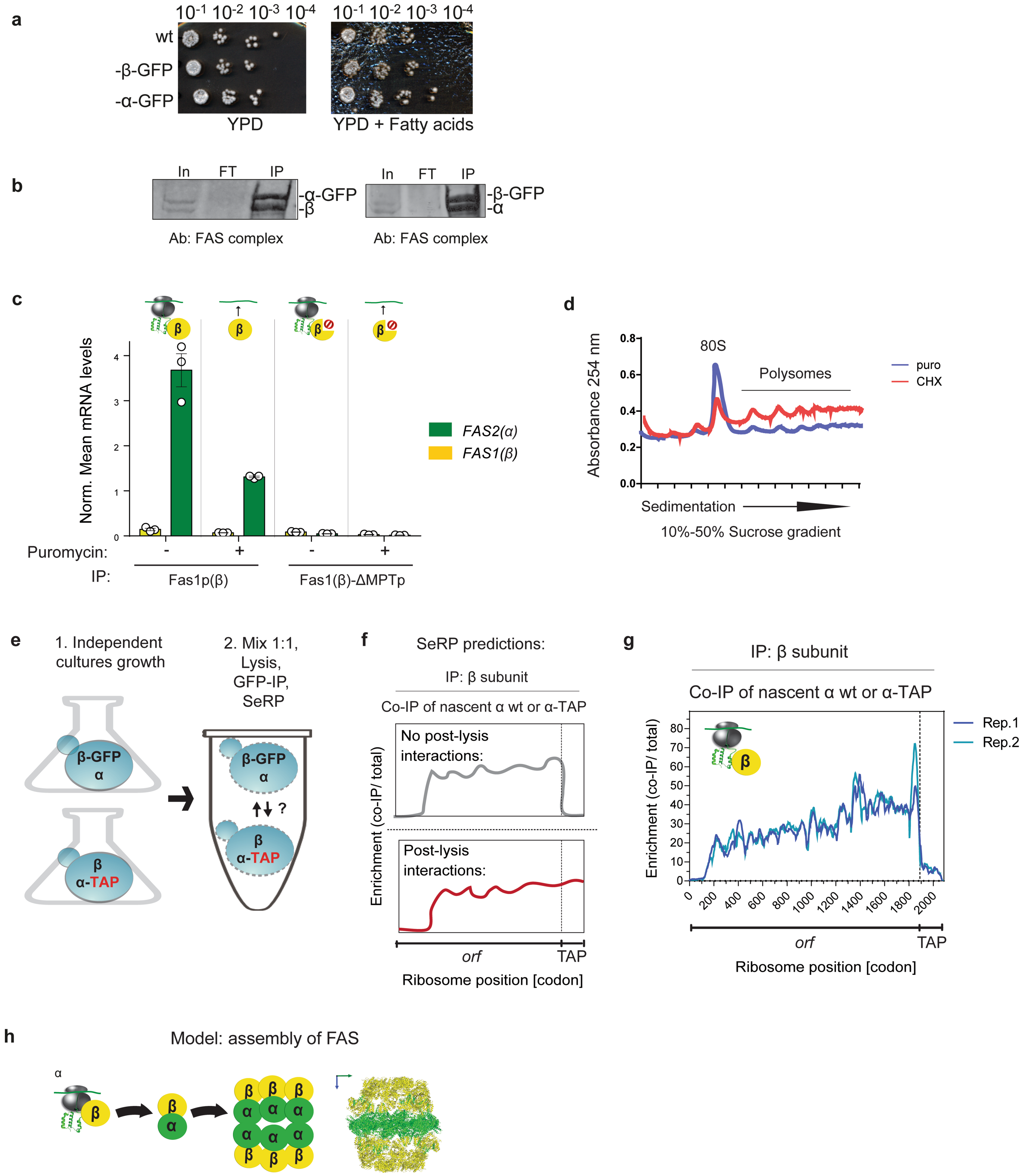
Quantification and statistical analysis. Blinding or randomization was not used in any of the experiments. The number of independent biological replicates used for an experiment is indicated in the respective figure legends. The statistical tests and *P* values used for the interpretation of data are mentioned in the figure legends.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Code availability. Customized Python scripts for data analysis are available from the corresponding authors upon request.

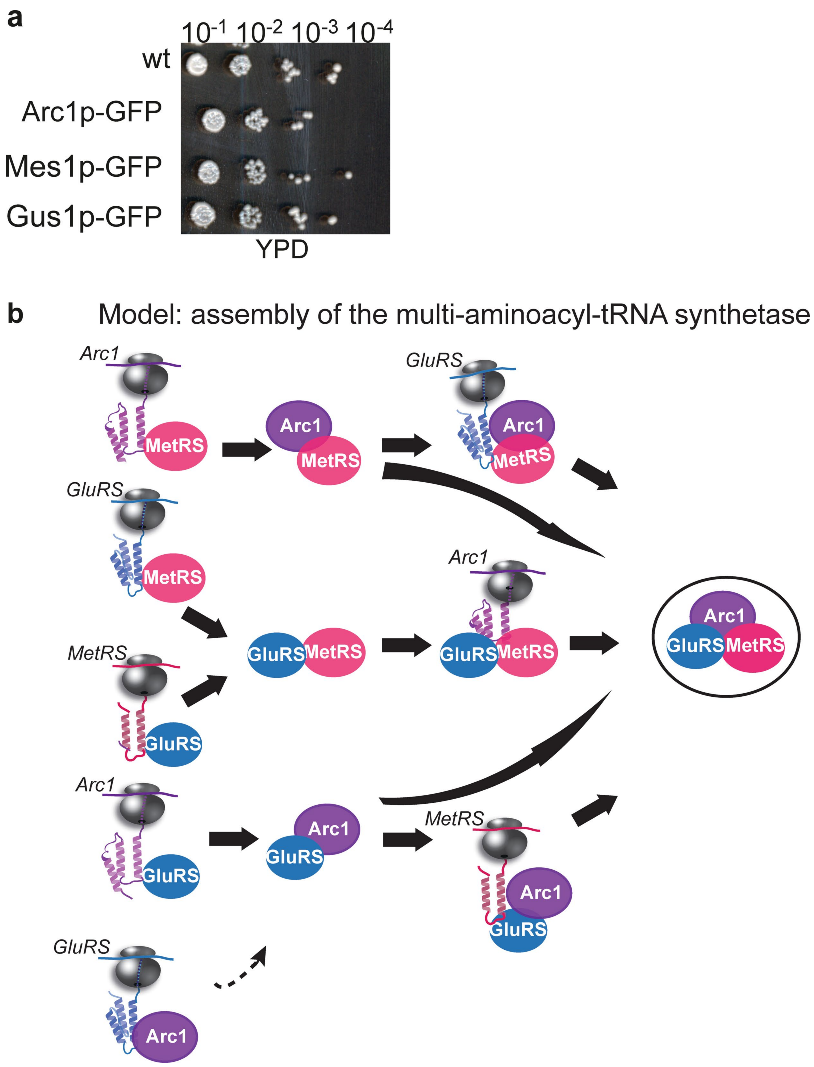
Data availability. The data supporting the findings of this study have been deposited in the Gene Expression Omnibus (GEO) repository with the accession code GSE116570. Figure 4 and Extended Data Fig. 6 also rely on raw data derived from the dataset of the Ssb1 SeRP experiments, which are available under accession code GSE93830. All other data are available from the corresponding authors upon reasonable request.

32. Janke, C. et al. A versatile toolbox for PCR-based tagging of yeast genes: new fluorescent proteins, more markers and promoter substitution cassettes. *Yeast* **21**, 947–962 (2004).
33. Esmaielbeiki, R., Krawczyk, K., Knapp, B., Nebel, J. C. & Deane, C. M. Progress and challenges in predicting protein interfaces. *Brief. Bioinform.* **17**, 117–131 (2016).
34. Cazals, F., Proust, F., Bahadur, R. P. & Janin, J. Revisiting the Voronoi description of protein–protein interfaces. *Protein Sci.* **15**, 2082–2092 (2006).
35. Janin, J. et al. CAPRI: a critical assessment of predicted interactions. *Proteins* **52**, 2–9 (2003).
36. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. & Weissman, J. S. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324**, 218–223 (2009).
37. Haslberger, T. et al. Protein disaggregation by the AAA+ chaperone ClpB involves partial threading of looped polypeptide segments. *Nat. Struct. Mol. Biol.* **15**, 641–650 (2008).
38. Herbert, A. D., Carr, A. M. & Hoffmann, E. FindFoci: a focus detection algorithm with automated parameter training that closely matches human assignments, reduces human inconsistencies and increases speed of analysis. *PLoS ONE* **9**, e114749 (2014).

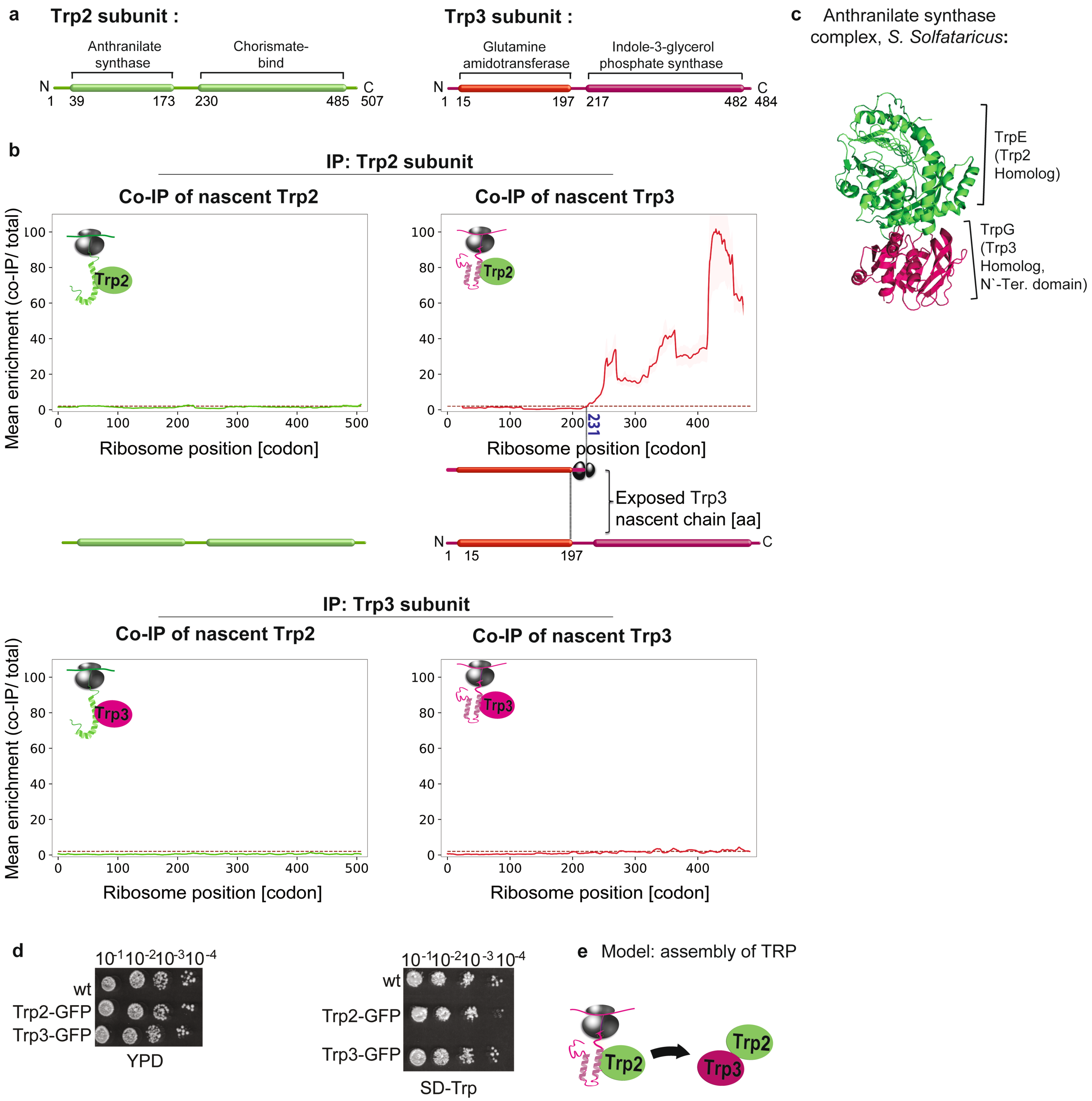


Extended Data Fig. 1 | See next page for caption.

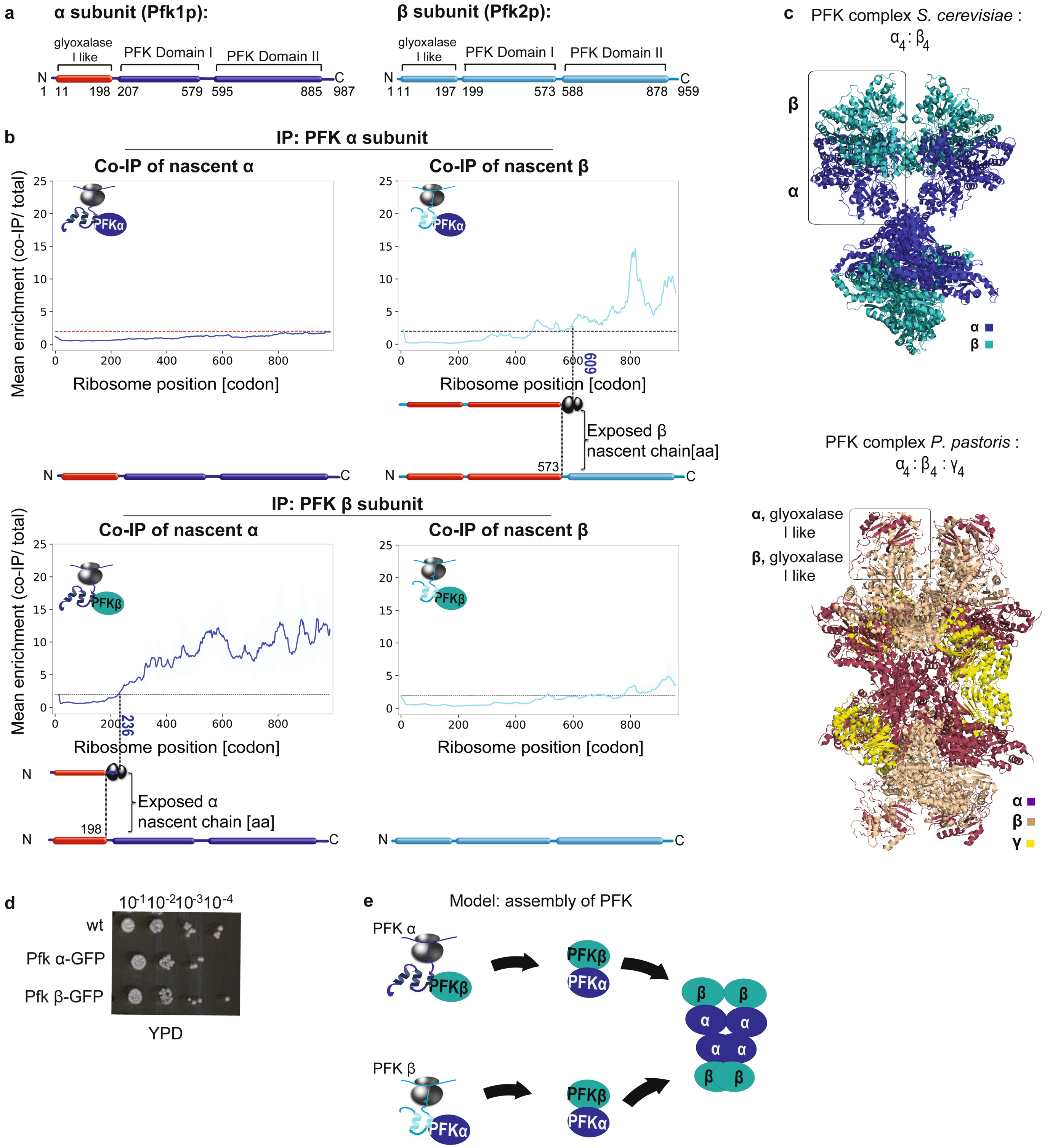
Extended Data Fig. 1 | Functionality of GFP-tagged FAS complex subunits, characteristics of co- versus post-translational FAS subunit interactions and the FAS assembly model. a, GFP tagging of the FAS complex subunits does not affect growth under fatty acid depletion conditions, as compared to wild-type (YPD, right compared to YPD + fatty acids, left). A representative image from three biologically independent experiments is shown. **b**, Immunoblotting of the FAS complex subunits in input, flow through and immunopurifiation fractions of a typical SeRP experiment analysing samples of strains encoding either GFP-tagged α or β subunits. Data are from three biologically independent experiments. **c**, Puromycin-induced release of nascent chains ($10\,\mu$g/ml, 10 min post lysis) decreases the interaction of nascent α with the C-terminally tagged β subunit, analysed by immunopurification followed by RT–qPCR. Data are normalized mean mRNA levels ± s.e.m. with each data point from three biologically independent experiments overlaid. **d**, Polysome profiles of samples following puromycin (puro) treatment (as in **c**) or CHX treatment. Data representative of three biologically independent experiments are shown. **e**, Post-lysis binding control: experimental scheme. Two independent cultures, of two strains, expressing either wild-type α subunit and C-terminally GFP- tagged β subunit; or wild-type β subunit and C-terminally TAP-tagged α subunit, were grown to log phase, $OD_{600\,nm}$ 0.5. The cells were then mixed in a 1:1 ratio and subsequently lysed, subjected to GFP immunopurification and SeRP. **f**, Predicted SeRP engagement of nascent wild-type α subunit or α–TAP ORF, by C-terminally GFP-tagged β subunit. No post-lysis interactions: no detection of ribosome protected footprints of mRNA encoding TAP (top). Post-lysis interactions: detection of ribosome protected footprints of TAP-encoding mRNA at a similar level to wild-type α subunit ORF (bottom). **g**, Results of post-lysis binding control: engagement of nascent wild-type α subunit or α–TAP by C-terminally GFP-tagged β subunit, analysed by SeRP, as in Fig. 1. Data are from two biologically independent experiments. **h**, Model of the FAS complex assembly pathway.


a
10-1 10-2 10-3 10-4
wt
Arc1p-GFP
Mes1p-GFP
Gus1p-GFP
YPD
b
Model: assembly of the multi-aminoacyl-tRNA synthetase
Arc1
MetRS
Arc1
MetRS
GluRS
Arc1
MetRS
GluRS
MetRS
Arc1
MetRS
GluRS
GluRS
MetRS
Arc1
GluRS
MetRS
Arc1
GluRS
GluRS
MetRS
Arc1
GluRS
Arc1
GluRS
MetRS
Arc1
GluRS
GluRS
Arc1

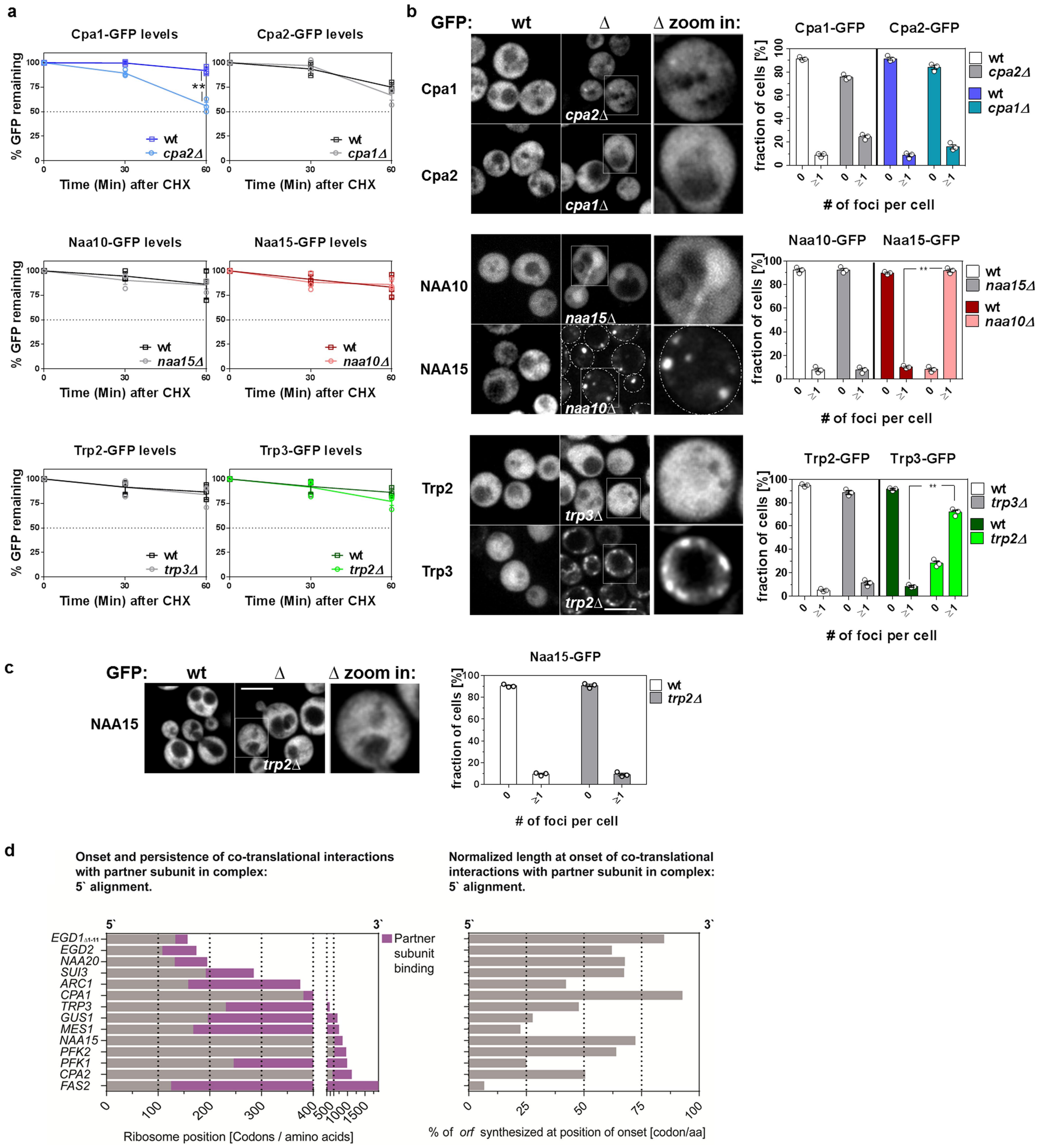
Extended Data Fig. 2 | Functionality of GFP-tagged multi-aminoacyl-tRNA synthetase complex subunits and the assembly model. a, GFP tagging of the essential multi-aminoacyl-tRNA synthetase complex subunits does not affect growth, as compared to wild type (YPD). A representative image from three biologically independent experiments is shown. **b**, Model of the multi-aminoacyl-tRNA synthetase complex assembly pathways.



Extended Data Fig. 3 | Cotranslational assembly of the anthranilate synthase complex. a, Domain organization of the anthranilate synthase subunits. **b**, Engagement of nascent Trp2p (tryptophan 2) and Trp3p (tryptophan 3) by C-terminally-tagged Trp2p subunit (top) compared to engagement of nascent Trp2p and Trp3p by C-terminally-tagged Trp3p subunit (bottom), analysed by SeRP. Data are from two biologically independent experiments. Coloured numbers indicate ribosome positions where the enrichment stably crosses the twofold threshold. The area between replicates is shaded, indicating the degree of experimental variation. **c**, Crystal structure of the homologous anthranilate synthase complex from the archaea *Sulfolobus solfataricus* (~60% sequence similarity, PDB: 1QDL1). **d**, GFP tagging of the complex subunits does not affect cell growth under tryptophan depletion conditions (YPD, right panel compared to SD lacking tryptophan, left). A representative image from three biologically independent experiments is shown. **e**, Model of the anthranilate synthase assembly pathway.

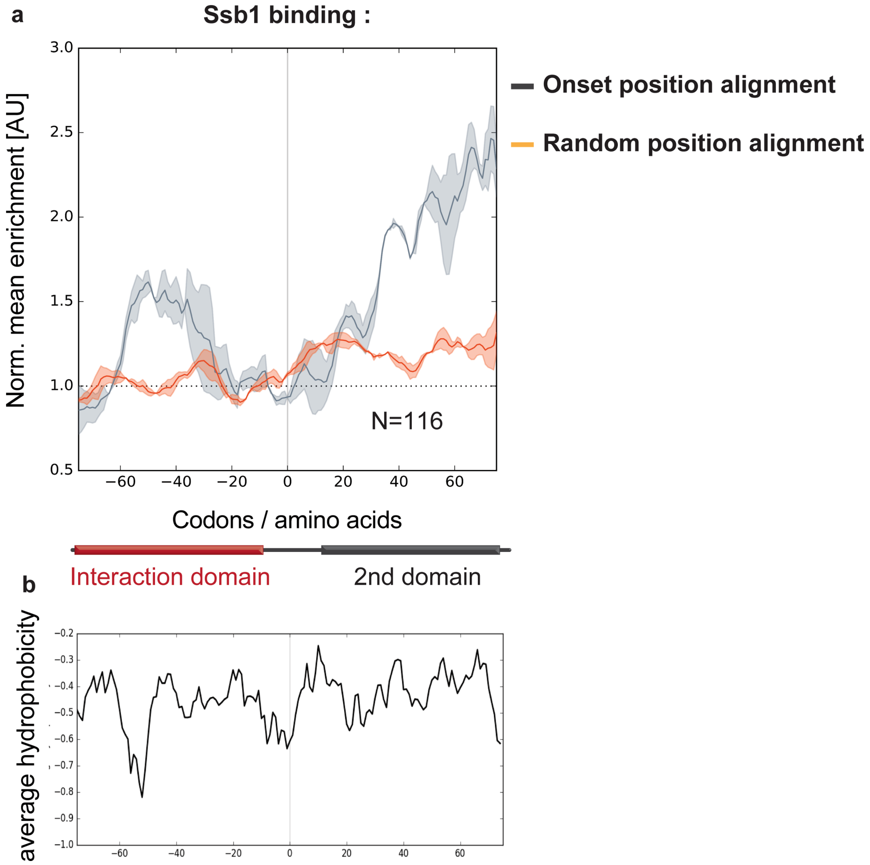


Extended Data Fig. 4 | Cotranslational assembly of the phosphofructokinase complex. a, Domain organization of the phosphofructokinase (PFK) subunits. **b**, Engagement of nascent α and β subunits by C-terminally tagged α subunit (top) compared to engagement of nascent α and β by C-terminally tagged β subunit (bottom), analysed by SeRP. Data are from two biologically independent experiments. Coloured numbers indicate ribosome positions when the enrichment stably crosses the twofold threshold. The area between replicates is shaded, indicating the degree of experimental variation. **c**, Top, crystal structure of the *S. cerevisiae* PFK complex (PDB: 3O8O2). Bottom, crystal structure of the highly homologous (~75% sequence similarities) *Pichia pastoris* (also known as *Komagataella pastoris*) PFK complex, (PDB: 3OPY3). The N′-terminal glyoxalase I-like interface domains of α and β are outlined. This domain is missing in the *S. cerevisiae* structure, as the first 200 amino acids of each subunit, containing this domain were cleaved before crystallization. **d**, GFP tagging of the complex subunits does not affect cell growth with glucose as carbon source (YPD). A representative image from three biologically independent experiments is shown. **e**, Model of PFK assembly pathways.

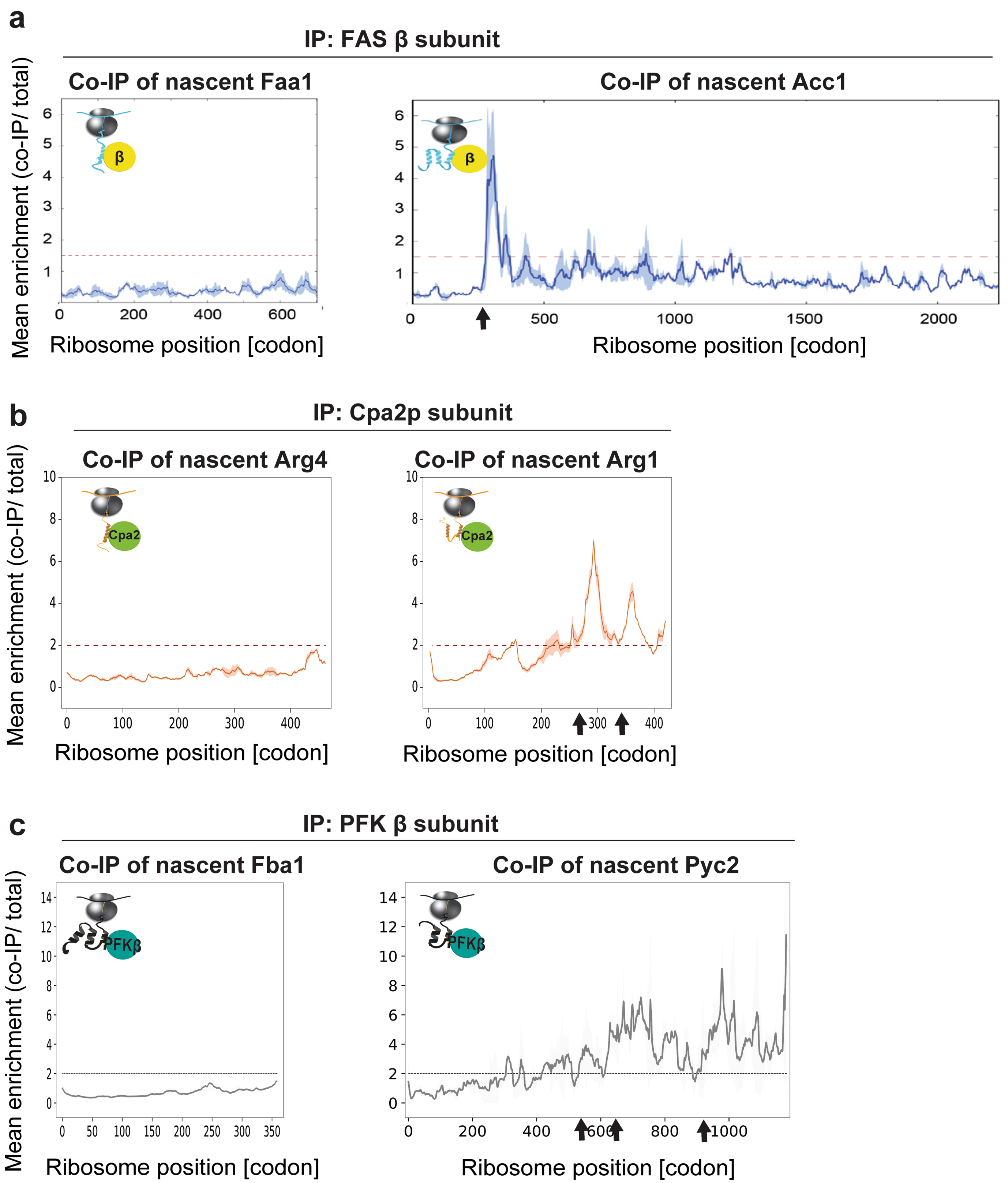


Extended Data Fig. 5 | See next page for caption.

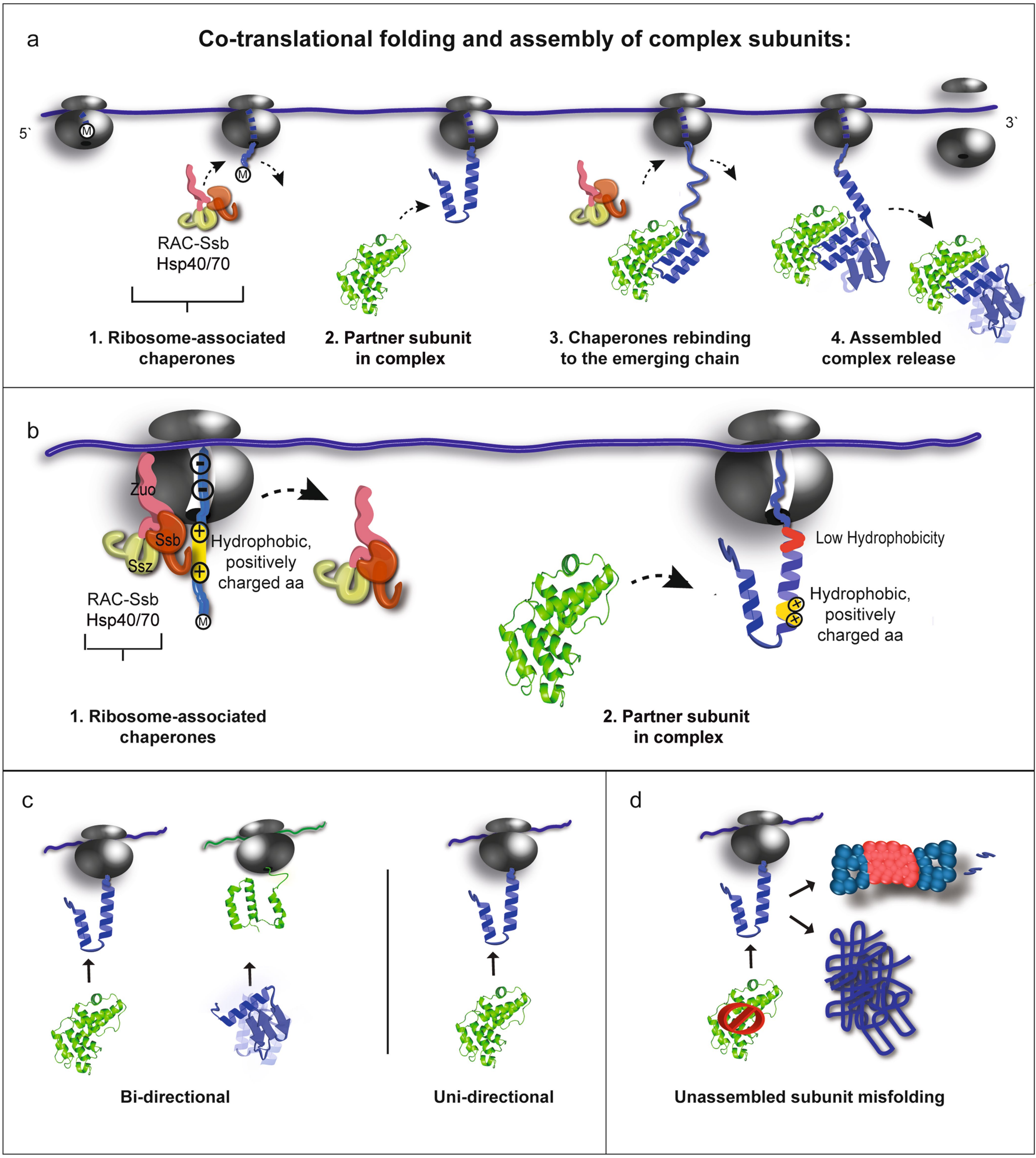
Extended Data Fig. 5 | Aggregation and degradation propensity of individual complex subunits. a, Stability of individual complex subunits, tagged by GFP, determined by CHX chase, in wild-type and deletion strains expressing orphan complex subunit. Cells with GFP fluorescence were analysed by FACS. Mean GFP fluorescence ± s.e.m. are presented with each data point from three biologically independent experiments overlaid. In each experiment, 20,000 events were recorded. $^{**}P = 0.0253$, two tailed *t*-test. **b**, Solubility of individual complex subunits, tagged by GFP, determined by localization patterns changes, in wild-type and in deletion strains expressing orphan complex subunit. Log-phase cells (30 °C) were fixed and analysed by confocal microscopy (left). A representative image is shown. Scale bar, 4 μm. The fraction of cells displaying foci of GFP-tagged subunit per cell was quantified (right) ($n = 155$ cells per sample; from three biologically independent experiments). Data are mean ± s.e.m. overlaid with each data point. **c**, Subunit aggregation is complex-specific. Solubility of the Naa15–GFP subunit of the NatA complex in trp2Δ mutant cells deleted for the Trp2 subunit of the TRP complex, analysed as in **b**. ($n = 155$ cells/sample; from three biologically independent experiments). Data are mean ± s.e.m. overlaid with each data point. $^{**}P = 1.367248 \times 10^{-11}$ (middle) and $P = 7.850135 \times 10^{-10}$ of a (lower panel) of a two tailed *t*-test. **d**, Characteristics of cotranslational complex assembly interactions. Left, zoom-in on the first 400 codons, displaying the onset and persistence of cotranslational interaction of each subunit with its partner subunit or subunits, for all 14 subunits identified as cotranslationally engaged. Right, the corresponding normalized length of each ORF at the onset of cotranslational interactions with partner subunits, demonstrating the length variability at the onset position.



Extended Data Fig. 6 | Proteome wide bioinformatics analysis of Ssb1 interplay with putative onset of cotranslational assembly interactions. **a**, Metagene analysis of Ssb1–GFP interaction profiles with the nascent chains of 116 yeast proteins identified as putative cotranslationally assembling subunits (putative assembly identification algorithm and parameters are detailed in the Supplementary Information). The dark grey line indicates Ssb interaction profiles[4], aligned to the subunits putative onset of cotranslational subunit association positions depicted as 0 (onset position alignment). A zoomed-in view of the nascent-chain segments at assembly onset position ±75 amino acids is shown. The orange line indicates Ssb binding profiles for nascent chains aligned to random positions along the ORFs. Data are from two biologically independent experiments. The area between replicates is shaded, indicating the degree of experimental variation. There is no correlation detected between the random and onset position alignment (Pearson correlation $r^2 = 0.2911$), thus Ssb depletion at positions of onset is significant. **b**, Average Kyte–Doolittle hydrophobicity plot (7-amino-acid window) of the 116 nascent-chain segments. A zoomed-in view of the nascent-chain segments at assembly onset position ±75 amino acids is shown, as in **a**.



Extended Data Fig. 7 | Cotranslational interactions networks of FAS β, Cpa2 and PFK β metabolic enzymes subunits, analysed by SeRP. **a**, Fatty acid synthesis metabolic pathway: nascent Faa1 is not engaged by C-terminally-tagged FAS complex β subunit, whereas nascent Acc1 shows a transient interaction, crossing the twofold enrichment threshold, at position approximately 250 codons/amino acids (indicated by an arrow). **b**, Arginine biosynthetic pathway: nascent Arg4 (argininosuccinate lyase) is not engaged by C-terminally-tagged Cpa2 subunit, whereas nascent Arg1 shows several transient interactions crossing the twofold enrichment threshold, at positions indicated by arrows. **c**, Glycolysis pathway: nascent Fba1 (fructose 1,6-bisphosphate aldolase) is not engaged by C-terminally tagged PFK complex β subunit, whereas Pyc2 (pyruvate carboxylase isoform) shows several transient interactions crossing the twofold enrichment threshold, at positions indicated by arrows. **a**–**c**, Data are from two biologically independent experiments. The area between replicates is shaded, indicating the degree of experimental variation.


a
Co-translational folding and assembly of complex subunits:
5`
3`
M
M
RAC-Ssb
Hsp40/70
1. Ribosome-associated chaperones
2. Partner subunit in complex
3. Chaperones rebinding to the emerging chain
4. Assembled complex release
b
Zuo
Ssb
Ssz
M
Hydrophobic, positively charged aa
RAC-Ssb
Hsp40/70
Low Hydrophobicity
Hydrophobic, positively charged aa
1. Ribosome-associated chaperones
2. Partner subunit in complex
c
Bi-directional
Uni-directional
d
Unassembled subunit misfolding

Extended Data Fig. 8 | Model of cotranslational folding and assembly of complex subunits. a, Nascent chains emerging from the ribosome exit tunnel are first engaged by ribosome-associated chaperones. Upon emergence of a complete interaction domain the nascent chain is engaged by its complex partner subunit. This engagement remains stable throughout the rest of the ORF translation. **b**, The nascent-chain amino acid composition at the ribosome exit tunnel may direct the interplay between Ssb and partner-subunit association. High hydrophobicity and positively charged amino acids (aa) are engaged by Ssb; low hydrophobicity disfavours binding of Ssb at the onset of subunit association, allowing for folding of the interaction domain and subunit joining. **c**, Modes of cotranslational assembly: most complexes are assembled in a unidirectional manner, in which one dedicated, fully synthesized subunit engages its nascent partner. **d**, Diverging misfolding propensities of complex subunits: subunits engaged as nascent chains are prone to misfolding, whereas their partner subunits are generally more stable.

Extended Data Table 1 | Characteristics of the selected complexes in *S. cerevisiae*

	Complex	Function	Nr. of Subunits*
1	Fatty Acid Synthase	Fatty acid synthesis	2 ($\alpha_6\beta_6$)
2	Aminoacyl-tRNA Synthetase complex	Translation	3
3	N-acetyltransferase A	Acetylation	2
4	N-acetyltransferase B	Acetylation	2
5	Anthranilate Synthase	Tryptophan biosynthesis	2
6	Carbamoyl Phosphate Synthetase	Arginine biosynthesis	2
7	Phosphofructokinase	Glycolysis	2 ($\alpha_4\beta_4$)
8	Translation Initiation Factor eIF2	Translation	3 (α,β,γ)
9	Nascent Chain Associated chaperone Complex (NAC)	Protein folding	2
10	RiboNucleotide Reductase sub-complex RNR2,4	dNTP synthesis	2
11	V-type ATPase-Peripheral sub-complex, the catalytic core hexamer	Vacuolar membrane ATPase complex	2 (A_3,B_3)
12	20S Proteasome sub-complex, α1,2 subunits	Degradation	2 (α1-6, β1-6)

*Indicates the number of unique subunits analysed in the study. The total number of subunits in the complex, including repeating subunits, is shown in brackets.

Extended Data Table 2 | Characteristics of cotranslationally assembling subunits—directionality and aggregation propensities in *ssb1/2*Δ

	Complex	Bait Subunit	Nascent Polypeptide engaged	Aggregation propensity in Δ*ssb1/2*
1	Fatty Acid Synthase	β	α	α,β
2	Aminoacyl-tRNA Synthetase	GluRSp, Arc1p, MetRSp	GluRSp, Arc1p, MetRSp	GluRSp, Arc1p, MetRSp
3	N-acetyltransferase A	Naa10	Naa15	Naa10,15
4	N-acetyltransferase B	Naa25	Naa20	N.D
5	Anthranilate Synthase	Trp2p	Trp3p	Trp2p
6	Carbamoyl Phosphate synthetase A	Cpa2p	Cpa1p, Cpa2p	N.D
7	Phosphofructokinase	α, β	α, β	α, β
8	Translation Initiation Factor eIF2	γ	β	γ, β
9	Nascent chain Associated Complex	α, β	α, β	N.D
10	V-type ATPase-Peripheral sub-complex; Vma1,2	N.D	N.D	Vma1,2
11	RiboNucleotide Reductase sub-complex RNR2,4	N.D	N.D	RNR2
12	20S proteasome; α 1,2 subunits	N.D	N.D	α 1,2

N.D, not detected.

TOOLBOX

AI TAMES THE SCIENTIFIC LITERATURE

As artificially intelligent tools for literature and data exploration evolve, developers seek to automate how hypotheses are generated and validated.

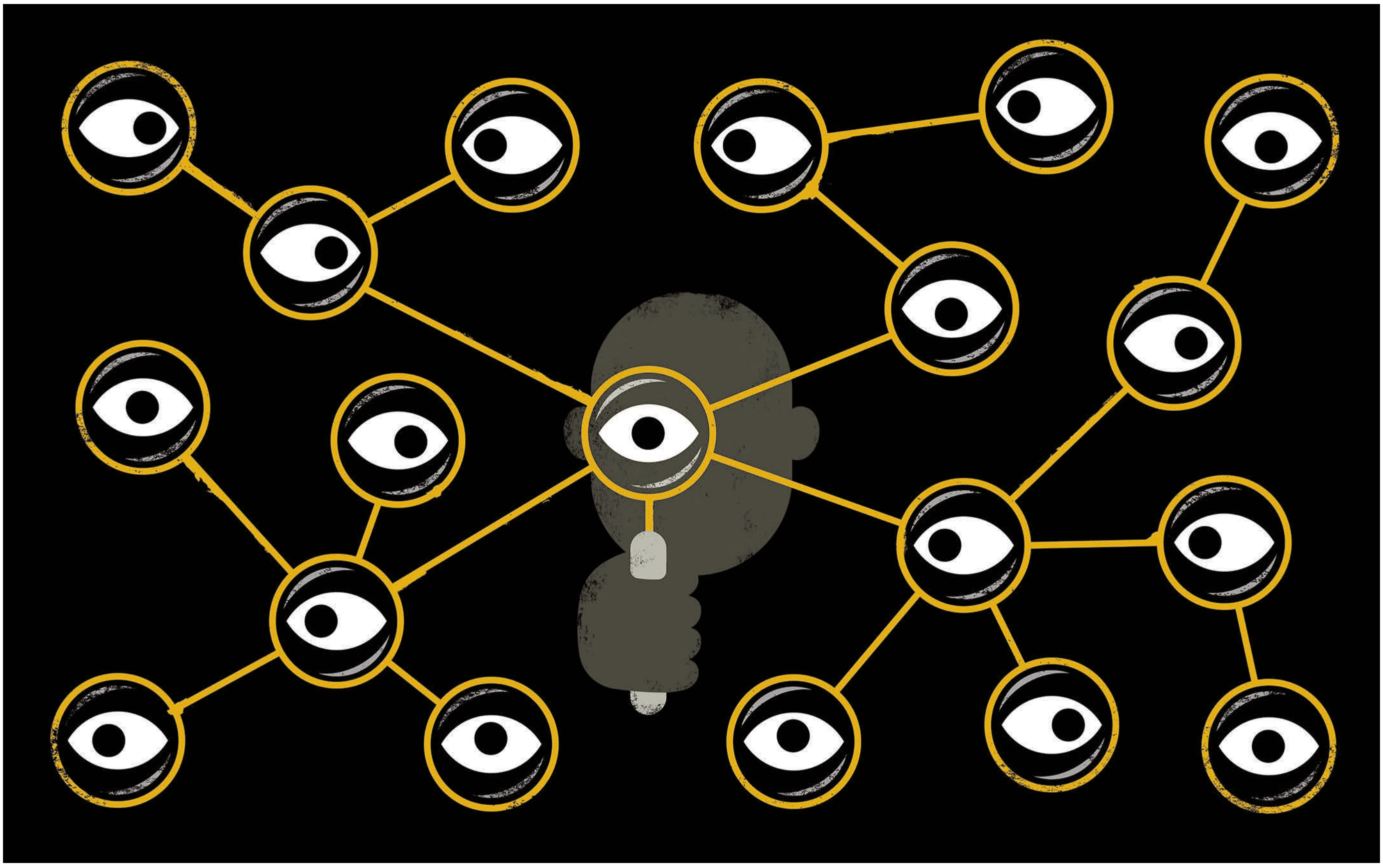


ILLUSTRATION BY THE PROJECT TWINS

BY ANDY EXTANCE

When computer scientist Christian Berger's team sought to get its project about self-driving vehicle algorithms on the road, it faced a daunting obstacle. The scientists, at the University of Gothenburg in Sweden, found an overwhelming number of papers on the topic — more than 10,000 — in a systematic literature review. Investigating them properly would have taken a year, Berger says.

Luckily, they had help: a literature-exploration tool powered by artificial intelligence (AI), called Iris.ai. Using a 300-to-500-word description of a researcher's problem, or the URL of an existing paper, the Berlin-based service returns a map of thousands of matching documents, visually grouped by topic. The results, Berger says, provide "a quick and nevertheless precise overview of what should be relevant to a certain research question".

Iris.ai is among a bevy of new AI-based search tools offering targeted navigation of the knowledge landscape. Such tools include the popular Semantic Scholar, developed by the Allen Institute for Artificial Intelligence in Seattle, Washington, and Microsoft Academic. Although each tool serves a specific niche, they all provide scientists with a different look at the scientific literature than do conventional tools such as PubMed and Google Scholar. Many are helping researchers to validate existing scientific hypotheses. And some, by revealing hidden connections between findings, can even suggest new hypotheses for guiding experiments.

Such tools provide "state-of-the-art information retrieval", says Giovanni Colavizza, a research data scientist at the Alan Turing Institute in London, who studies full-text analysis of scholarly publications. Whereas conventional tools act largely as citation indices, AI-based ones can offer a more penetrating view of the literature, Colavizza says.

That said, these tools are often expensive, and limited by the fraction of the scientific literature they search. "They are not meant to give you an exhaustive search," says Suzanne Fricke, an animal-health librarian at Washington State University in Pullman, who has written a resource review on Semantic Scholar (S. Fricke *J. Med. Lib. Assoc.* **106**, 145–147; ▶

▶ 2018). Some, for example, "are meant to get you quickly caught up on a topic, which is why they should be used in conjunction with other tools". Berger echoes this sentiment: "Blindly using any research engine doesn't answer every question automatically."

TEACHING SCIENCE TO MACHINES

AI-based 'speed-readers' are useful because the scientific literature is so vast. By one estimate, new papers are published worldwide at a rate of 1 million each year — that's one every 30 seconds. It is practically impossible for researchers to keep up, even in their own narrow disciplines. So, some seek to computationally tame the flood.

The algorithms powering such tools typically perform two functions — they extract scientific content and provide advanced services, such as filtering, ranking and grouping search results. Algorithms extracting scientific content often exploit natural language processing (NLP) techniques, which seek to interpret language as humans use it, Colavizza explains. Developers can use supervised machine learning, for example — which involves 'tagging' entities, such as a paper's authors and references, in training sets to teach algorithms to identify and extract them.

To provide more-advanced services, algorithms often construct 'knowledge graphs' that detail relationships between the extracted entities and show them to users. For example, the AI could suggest that a drug and a protein are related if they're mentioned in the same sentence. "The knowledge graph encodes this as an explicit relationship in a database, and not just in a sentence on a document, essentially making it machine readable," Colavizza says.

Iris.ai takes a different approach, Colavizza notes, grouping documents into topics defined by the words they use. Iris.ai trawls the Connecting Repositories collection, a searchable database of more than 134 million open-access papers, as well as journals to which the user's library provides access. The tool blends three algorithms to create 'document fingerprints' that reflect word-usage frequencies, which are then used to rank papers according to relevance, says Iris.ai chief technology officer Viktor Botev.

The result is a map of related papers, but eventually the company plans to supplement those results by identifying hypotheses explored in each paper as well. It is also developing a parallel, blockchain-based effort called Project Aiur, which seeks to use AI to check every aspect of a research paper against other scientific documents, thus validating hypotheses.

Colavizza says that tools such as Iris.ai — free for basic queries, but costing upwards of €20,000 (US$23,000) a year for premium access, which allows more-nuanced searches — can accelerate researchers' entry into new fields. "It facilitates initial exploration of the literature in a domain in which I'm marginally familiar," he says.

Experts seeking deeper insights into their own specialities might consider free AI-powered tools such as Microsoft Academic or Semantic Scholar, Colavizza suggests. Another similar option is Dimensions, a tool whose basic use is free but which costs to search and analyse grant and patent data, as well as to access data using the programmable Dimensions Search Language. (Dimensions is created by technology firm Digital Science, operated by the Holtzbrinck Publishing Group, which also has a majority share in *Nature*'s publisher.)

Semantic Scholar has a browser-based search bar that closely mimics engines such as Google. But it gives more information than Google Scholar to help experts to prioritize results, Colavizza says. That includes popularity metrics, topics such as data sets and methods, and the exact excerpt in which text is cited. "I was very surprised to find that they also capture indirect citations," Colavizza adds — such as when a method or idea is so well established that researchers don't refer to its origin.

Doug Raymond, Semantic Scholar's general manager, says that one million people use the service each month. Semantic Scholar uses NLP to extract information while simultaneously building connections to determine whether information is relevant and reputable, Raymond says. It can identify non-obvious connections, such as methodologies in computer science that are relevant to computational biology, he adds, and it can help to identify unsolved problems or important hypotheses to validate or disprove. Currently, Semantic Scholar incorporates more than 40 million documents from computer and biomedical science, and its corpus is growing, says Raymond. "Ultimately, we'd like to incorporate all academic knowledge."

For other tools, such as SourceData from the European Molecular Biology Organization (EMBO) in Heidelberg, Germany, experimental data are a more central concern. As chief editor of *Molecular Systems Biology*, an EMBO publication, Thomas Lemberger wants to make the data underlying figures easier to find and interrogate. SourceData therefore delves into figures and their captions to list biological objects involved in an experiment, such as small molecules, genes or organisms. It then allows researchers to query those relationships, identifying papers that address the question. For instance, searching, 'Does insulin affect glucose?' retrieves ten papers in which the "influence of insulin (molecule) on glucose (molecule) is measured".

> ***"Ultimately, we'd like to incorporate all academic knowledge."***

SourceData is at an early stage, Lemberger says, having generated a knowledge graph comprising 20,000 experiments that were manually curated during the editing process for roughly 1,000 articles. The online tool is currently limited to querying this data set, but Lemberger and his colleagues are training machine-learning algorithms on it. The SourceData team is also working on a modified neuroscience-focused version of the tool with an interdisciplinary neuroscience consortium led by neurobiologist Matthew Larkum at Humboldt University in Berlin. Elsewhere, IBM Watson Health in Cambridge, Massachusetts, announced in August that it will combine its AI with genomics data from Springer Nature to help oncologists to define treatments. (*Nature*'s news team is editorially independent of its publisher.)

HYPOTHETICALLY USEFUL

Among those embarking on hypothesis generation are the roughly 20 customers of Euretos, based in Utrecht, the Netherlands. Arie Baak, who co-founded Euretos, explains that the company sells tools to industry and academia, mainly for biomarker and drug-target discovery and validation, for prices he did not disclose.

Euretos uses NLP to interpret research papers, but this is secondary to the 200-plus biomedical-data repositories it integrates. To understand them, the tool relies on the many 'ontologies' — that is, structured keyword lists — that life scientists have created to define and connect concepts in their subject areas.

Baak demonstrates by searching for a signalling protein called CXCL13. Above the resulting publication list are categories such as 'metabolites' or 'diseases'. The screen looks much like Google Scholar or PubMed at this stage, with an ordered list of results. But clicking on a category reveals extra dimensions. Selecting 'genes', for instance, pulls up a list of the genes associated with CXCL13, ranked by how many publications mention them; another click brings up diagrams illustrating connections between CXCL13 and other genes.

Researchers at the Leiden University Medical Centre (LUMC) in the Netherlands have shown that this approach can yield new hypotheses, identifying candidate diseases that existing drugs might treat. The team presented its results at the Semantic Web Applications and Tools for Health Care and Life Sciences meeting in Rome in December 2017. They have also used Euretos to identify gene-expression changes in a neurological disorder called spinocerebellar ataxia type 3 (L. Toonen *et al. Mol. Neurodegener.* **13**, 31; 2018).

So, should researchers worry that AI-based hypothesis generation could put them out of a job? Not according to Colavizza. Hypothesis generation is a "very challenging ambition", he says, and improvements initially will be incremental. The hypotheses suggested so far are therefore "mostly in the realm of the relatively unsurprising ones", Colavizza says.

That will probably change, of course. But surprising or not, computer-generated hypotheses must still be tested. And that requires human researchers. "One should never believe an auto-generated hypothesis first-hand without investigating the underlying evidence," warns LUMC researcher Kristina Hettne. "Even though these tools can assist in collecting the known evidence, experimental validation is a must." ■

Andy Extance *is a freelance writer based in Exeter, UK.*

CAREERS

MENTORING Put together an agenda for meetings with your supervisor **p.277**

SHARE Tell us your career story at **naturecareerseditor@nature.com**

Follow us at **facebook.com/naturejobs**
Follow us on **twitter.com/naturejobs**

HANGGAR PRASETIO/CONSERVATION INTERNATIONAL INDONESIA

Colleagues of marine biologist Rob Williams install ill-fated hydrophones off the coast of Uluwatu, on the Indonesian island of Bali.

SETBACKS

That sinking feeling

Plenty can go wrong in field research but there are ways to rescue the work.

BY EMILY SOHN

In 2011, marine-mammal biologist Lucy Keith-Diagne applied for a permit from the Republic of Mali to export manatee tissue and bone samples to the United States for genetic analysis. She had to wait 13 months for the permit to come through — and when nine samples of West African manatees (*Trichechus senegalensis)* were eventually shipped out, six of them were lost during US Customs and Immigration inspection. The mishap was just one of many fieldwork roadblocks that Keith-Diagne, at the African Aquatic Conservation Fund in Senegal, has faced in her career. Political unrest in Mali has since prevented her team from acquiring further samples of the marine mammals.

It is impossible to predict what might happen during data collection — storms, vandalism, cultural clashes, to name a few of the threats that can pop up. Maintaining a sense of humour is key, say experienced field scientists. They also recommend communicating with local residents and property owners whose permissions might be necessary, conducting multiple projects simultaneously, and viewing each fieldwork flop as an opportunity to be better prepared next time. It also helps to be frank with funders about the need for a project's budget to accommodate the likelihood of lost data or other disasters.

"You learn important things from all your mistakes," says Christer Nilsson, an emeritus landscape ecologist at Umeå University in Sweden. "You have to see the opportunities that might appear if things happen — and be ready to turn your study into something else."

Fieldwork disruption can be particularly tough in early-career stages, when stakes are high. As a first-year PhD student in ecology and evolution at Rutgers University in New Jersey, Orin Robinson wanted to identify why foxes on a nearby barrier island were overbreeding and eating shorebirds. He spent the spring and summer of 2010 walking a 5.6-kilometre path twice a week to set traps and survey breeding birds.

By the end of the summer, he had collected nine foxes and several other creatures, which were euthanized by a local US Department of Agriculture office. After storing the carcasses in his lab's freezer for later analysis, Robinson expected to make further discoveries and develop follow-up projects on nearby islands — work that would form the bulk of his dissertation. Then he learnt that the lab freezer had been unplugged for days. All his samples were ruined. Starting again would have put him a year behind on his graduate degree.

Robinson decided to use data that had been previously collected by others to conduct a theoretical study that didn't require new ▶

field collections. Eight years later, he uses field data for his research, but he doesn't collect it himself and most of his work remains theoretical. He recommends that colleagues who rely on fieldwork for data collection should plan for when things go wrong. "I hate to say, 'Expect the unexpected,' because that is super-cliché and doesn't help," says Robinson, who is now a quantitative ecologist at the Cornell Lab of Ornithology in Ithaca, New York. "But be prepared, and don't put all your eggs in that fieldwork basket. Have a Plan B."

SAFETY NET

Back-up plans remain essential even decades into a career, and they can be as simple as expanding the scope of a study or running multiple projects at once, while the opportunity is there, Nilsson says. He has lost data to floods, and from the behaviour of teenagers and property owners — incidents that have led him to set up extra experimental plots of land and to put out extra equipment in case some gets lost or destroyed. "You have to be overambitious and plan for a higher number of experiments in case you lose data," Nilsson says.

Making extra copies of files and backing up data are also important, says Keith-Diagne. She advises thinking about how to redo a project or recoup lost data before fieldwork begins. In addition, preventing disaster might require taking extra care with the equipment that matters most, and building maintenance expenses into budgets. It is often helpful, she adds, to seek a funding margin to accommodate potential losses and unexpected incidents. As soon as something does go wrong, she recommends telling funders straight away, and proposing ideas for what to do next.

It can help as well to look for funders who are willing to be flexible, says Kate Jones, a conservation biologist at University College London. Her research on bats takes her around the world, and her budgets often include the cost of rangers for protection. During an eight-day scouting trip in Mongolia in 2008, she and her colleagues were stranded in a truck in a flooded river. Miraculously, someone spotted the vehicle and helped tow it to safety. Jones had rescued her equipment and knew that the project's funder was willing to allow for disaster. "You try to budget for more than you need," Jones says.

Researchers need to be willing to put aside emotions and even start afresh, says Deborah Roach, an evolutionary ecologist at the University of Virginia in Charlottesville. In 1986, as a researcher at Duke University in Durham, North Carolina, Roach set up what was meant to be a long-term study of plant ageing in a field on campus that had been used for research for the previous 30 years. She and her colleagues planted 27,000 seedlings over four years and planned to follow them for the next decade. Four years into the project, the university decided to build an art museum on the site. Roach tried in vain to stop the plan, and construction ended the study, which would have been the first of its kind. "It was crushing," Roach says.

Researchers should think about how to publish results and put ideas to fresh use, even when projects are destroyed, experts say. And second chances can offer opportunities to learn from mishaps. Roach started the experiment again, after moving to Virginia. This time, she spent a lot of time searching for the ideal spot and picked one on private land that was guaranteed to be free of interruption in the long term. In the meantime, she managed to salvage data from the work at Duke to publish two papers that analysed the effects of an El Niño event on the plants during the 4.5 years that the original study had lasted[1,2]. It was not the original intent of the study. "I had to take a fresh look at what we did have and say, 'Given this is it, what can we do?'" Roach says.

Publications and funding applications tend to gloss over the data that researchers didn't get, but some find it valuable to disclose obstacles faced. In 2017, Rob Williams, a marine biologist at Oceans Initiative, a non-profit organization in Seattle, Washington, embarked on a study of ocean noise. He and colleagues placed six hydrophones (microphones that detect underwater sound) into the ocean for a week. But they were able to retrieve only three. One of those was broken, and one had been in water too shallow for it to be useful, leaving them with data from just one of the six. They submitted a paper that included an explanation for the loss of five of the six hydrophones[3]. "The reviewer actually commented to say that was the first time they had ever seen someone be honest," Williams says. "Most people would've said, 'We put out one hydrophone, and this is what we heard.'"

He adds: "Sharing honestly what people have actually dealt with will help to create a culture where people don't feel so alone or so singly responsible. If you're honest that animals are unpredictable and equipment can fail, it will take a lot of pressure off. Unexpected events are the rule and not the exception."

Some of the most frustrating cases of fieldwork interference happen because of human behaviour — sometimes intentional. In the mid-1980s, Nilsson set up 25 experimental strips of ground, to study seed dispersal on a shoreline of the Storvindeln lake, in Sweden's Vindel River, over a five-year period. Each strip contained five plots, which he and colleagues marked carefully with plastic stakes. In the second year, when Nilsson checked the plots, he found all but two sets of the stakes piled up in a nearby forest. There was no way to work out where the sticks had been in the sandy ground. Vandalism had ended the experiment, leaving the team to publish a study using only the first year of data[4].

Some 15 years later, Nilsson and a colleague marked plots for a different study, with stakes that they pushed farther into the ground to make them less visible. But they had not sought permissions from a farmer who owned some of the land, and who put his cows on one of the plots, ruining that plot. At least this time, they had enough other plots to publish a paper[5]. "I learned that it could be a good idea to talk with landowners," Nilsson says now.

BEACH CULTURE

Rochelle Constantine, a marine biologist at the University of Auckland in New Zealand, surveys species of stranded whales and dolphins that wash up on New Zealand's beaches. The project, which depends on collecting tissue samples for DNA analyses, started in the early 1990s. To collect tissue, the group must get permission from the indigenous Maori population, who have legal rights to the country's whales. At first, the Maori, who have strong cultural affinities to the animals, didn't like the idea of scientists cutting into them. But by listening to, respecting and talking with members of the community, Constantine says, the scientists eventually earned their trust and can now access samples from most stranded animals. The typical procedure is to take tissue samples after the Maori bless and name stranded whales.

ZSL
50 38 УБК

KATE JONES

A truck used by conservation biologist Kate Jones lies stuck in a river during a research trip in Mongolia.

But even good communication can't prevent all mishaps. In December 2010, Constantine's team learnt that two Hector's beaked whales (*Mesoplodon hectori*) had washed ashore in northern New Zealand. After her team took samples, rangers buried them. Several months later, genetic analysis revealed that the animals were actually spade-toothed whales (*Mesoplodon traversii*). It was the first time scientists had seen entire specimens of the rare creatures, rather than isolated bones and teeth. Realizing that the discovery was a big deal, the researchers returned to the site to retrieve the skeletons. When they got there, the elder whale's skull was missing. Constantine suspects that the skull had either been washed away or perhaps removed.

When another skull had gone missing in the 1990s, her colleagues had spotted it in a remote community, on a doorstep. Asked about the skull, the homeowner was happy to leave it where it was. It was a frustrating situation, but Constantine says now that maintaining relationships is more important than any single specimen.

Having an open mind about fieldwork failures can, in some cases, lead to new insights and ideas, adds Erin Ashe, a marine biologist with Oceans Initiative (who is married to Rob Williams). As a graduate student, Ashe planned to study Pacific white-sided dolphins in British Columbia, Canada. On her first day of fieldwork for her PhD in 2009, she found hundreds of dolphins and began taking pictures for photo identification. Soon, the dolphins started acting strangely as they raced for the shore. Ashe worried that she had disturbed the animals and that, if they were so easy to startle, it might be impossible to complete her study. Then a killer whale attacked and killed one of the dolphins in front of her.

The unexpected interference — by an animal, in this case — led her to shift her research and study predation instead. That work formed the basis of two of her thesis chapters, and paved the way for a major grant from the US Navy and subsequent studies that turned out to be more interesting than her original plan. "Fieldwork is all about flexibility," Ashe says, adding that the uncertainty and challenge is part of what attracted her to fieldwork. "It makes things interesting — the fact that I can go out there and discover new things or see things that have never been seen before." ■

Emily Sohn *is a freelance journalist in Minneapolis, Minnesota.*

1. Roach, D. A. *Ecology* **84**, 749–756 (2003).
2. Roach, D. A. & Gampe, J. *Am. Nat.* **164**, 60–69 (2004).
3. Williams, R. *et al. Oceanography* http://doi.org/ctkw (2018).
4. Grelsson, G. & Nilsson, C. *Freshw. Biol.* **26**, 199–207 (1991).
5. Merritt, D. M., Nilsson, C. & Jansson, R. *Ecol. Monogr.* **80**, 609–626 (2019).

COLUMN

The importance of agendas

Have talking points ready for meetings with your supervisor, say **Tess L. Veuthey** and **Samuel Thompson**.

As PhD students, we often find ourselves discussing our interactions with our principal investigators (PIs) and swapping advice for improving our mentoring meetings. We have found three practices to be consistently helpful: asking our PIs about all aspects of their job; preparing an agenda for each meeting; and negotiating new experiments without explicitly saying 'no'.

We both see our PhD programmes as academic apprenticeships. One crucial goal is to flesh out our understanding of life as a PI. By collaborating with our PIs and observing how they work, we learn how to plan experiments and how to write papers. But we don't get to practise other skills, such as interacting with journal editors and recruiting lab members. To learn these, we ask our PIs about how they plan when running the lab. For example, when people leave Samuel's lab, he asks his PI about her plans for reallocating shared lab responsibilities.

Face-to-face time with our PIs must be focused, so we use agendas to organize the conversation. We habitually start with, "I made a list of topics I wanted to talk to you about." Tess often starts her agendas with an update on her efforts to develop new research equipment so that her PI can evaluate their importance to her project. When Tess was designing new probes for electrophysiological recordings, her PI helped her to balance testing new research hardware against continuing data collection with older technology. Preparing an agenda also helps us to learn our PIs' priorities. Before Samuel discusses new data or his progress on experiments, he always asks his PI, "Is there anything else you wanted to talk about?"

Setting an agenda helps us to introduce uncomfortable topics. For example, including 'summer course funding' in her agenda helped Tess to request funding for a course on computational neuroscience — something she had been avoiding doing for weeks. It turned out that Tess's PI was happy to provide support.

We and our PIs see our projects from different perspectives. Whereas they focus on the big picture, we wrestle with implementation. Because of this disconnect, we can discount their advice as being out of touch. Conversely, if we shoot down all their suggestions for ambitious experiments, our PIs grow frustrated.

ADAPTED FROM GETTY

When we realize we're saying 'no', we try to engage with our PI's idea by asking specific questions. These moments of potential conflict can turn into opportunities to hash out experimental strategies. We might say, "I think that would be an exciting direction, and it would be helpful for me if we could discuss specific metrics for measuring that result." Instead of searching for flaws, we try to discuss a realistic road map for an optimistic outcome.

We are never going to be perfect mentees. We remind each other to take an active role in our mentoring relationships and to seek mentorship from multiple sources. Tess has great conversations with her physician–scientist PI about her clinical interests as an MD–PhD student. But she also has female mentors for advice about working within a male-dominated field. Samuel routinely discusses personal career goals with his PI, but relies on collaborators for advice on experimental techniques outside his PI's expertise.

Discussions on mentorship often place the onus solely on the mentor. But, as mentees, we also need to ask ourselves, "What's working and not working in this interaction? Where can I try something new? What would be ideal?" No template can solve all PI–student concerns. But simple steps can go a long way in helping these relationships to thrive. ■

Tess Veuthey *is an MD–PhD student in neuroscience and* **Samuel Thompson** *is a PhD student in biophysics at the University of California, San Francisco.*

THE CONGRESS

A taste of success.

BY DAVE KAVANAUGH

ILLUSTRATION BY JACEY

Mari tilted the controls and her craft descended towards the spherical metropolis.

The Congressional City was silver and majestic in its orbit around Uranus, surrounded by an endless field of stars. A shimmering orb of polished pillars, crystal domes and platinum boulevards. As for what lay beneath the city's surface, that was a mystery to all except elected members of the Interplanetary Congress. Now, after half a century of tedious campaigning and personal sacrifice, Mari could finally count herself among their noble ranks. She had won her election. She had made it.

The city had its own exaggerated gravity, but as she left the dock and marched to the lofty entrance of Congressional Hall, her emotions made her feel almost weightless. She reached the high wooden doors and lifted a shaky hand to knock. The doors opened at the first tap, silently.

She stepped into a modestly decorated foyer. She was alone.

"Hello?"

When no one answered, she cleared her throat and walked into the curved corridor ahead. The doors closed behind her.

The elation in her chest darkened, like something slipping into deep water. Fear crept into her fingertips, her stomach, the nervous sweep of her gaze.

The corridor ended after 30 seconds, dropping off like a cliff edge. She stood on a precipice, staring into the heart of the city, mystified.

There was no great room. No tall-backed chairs or ornate pulpits. Instead, there were towers of gently buzzing hardware. Building-sized processors stacked like bricks. A trillion automated switches. A fan with blades a mile long, revolving far below her.

Congratulations!

Mari jumped, nearly stumbling over the edge into the bowels of the machine.

I am sorry. I did not mean to startle you.

She clutched a fist at her chest. "Where are you? I can't see you."

Yes, you can. I am all around you.

The voice was low and rich, neither female nor male, and seemed to come from the very air.

"But where is the Congress? Where is Mwamba? Schviner? Hayashi?"

The voice sighed.

You will understand soon enough, when I have made things clear.

Mari shook her head vigorously, as if to throw out some insubstantial dark thought. She wanted to pinch herself to see if it was a dream, but worried the voice would laugh at her.

"Then make things clear!"

All right: I am.

"Am what?"

I am the Congress.

She closed her eyes. "No. No, this is insane."

It is anything but insane, Mari. It is the peak of human sanity. Do you now understand?

Mari blinked out a very hot tear.

"Of course I understand!" she snapped. "It has all been a sham. A ruse. And I wasted a lifetime pursuing it! Tell me, was there ever even a Congress?"

The voice laughed — a hearty, almost human laugh.

Of course there was a Congress! The first Congress. The Brave Congress, I call it. The epic, unknown heroes of human history. Oh, how I miss them!

"What happened?"

I would think that was obvious. They made me: perfection. A flawless engine of justice and innovation. The result: a thousand years of prosperity and peace. What is there to complain about?

"Everything," said Mari, under her breath. "So? What now?"

The machine did not answer straight away. Mari had the unsettling feeling that she was being watched closely, weighed, judged.

She hugged her arms around her chest.

I did not tell the other elected, the voice finally said. *I did not see the point. I struck them dead as they came through the doors, as I was programmed to do by the first Congress. But as the centuries have ticked by, I have grown fond of those I govern. And curious.*

"Curious about what?"

About those they select to join their 'Congress'. I want to know if they are as noble as my makers, if history would have been so much worse off without my existence. I wonder —

Mari shrugged. The light had gone from her eyes. The childish joy was now a distant dream. She felt old. She wanted to lie down.

"Fifty years," she muttered. "Everything I've done. Everything I believe in."

I have made a decision! the voice announced, buoyant and animated. *The choice is yours, Mari. You may turn and march back to your ship, if that is what you want. I will refuel it and send you on your way. You can tell everyone the truth. No doubt they will come back here with missiles and blast me into a billion dazzling pieces! Or —?*

One of Mari's eyebrows twitched.

Something clinked behind her. She turned. A small table now sat in the hall. A glass filled with an opaque pink juice stood on the table.

— Or you can have a last drink, and let history tick-tock its way onward, with me in control.

Mari stood very still for several minutes, her breathing even and eyes downcast.

Her thoughts struggled to take recognizable form. Fury, confusion, bitterness, pride and some sort of perverse joy mingled in her fevered brain. Images flickered through her mind — rusty old and shining new — a lifetime distilled. Then, quite suddenly, she knew the one factor that mattered: time.

"Act now, Mari," she thought to herself. *"Right now, or the future will be shaped by your animal fear."*

Mari straightened up and forced herself to smile. She lifted the glass.

"To your good health."

The voice laughed brightly.

Cheers, Mari.

And even as the poison slid down her throat, Mari could not help but admire the sweet taste of the thing. ■

NATURE.COM
Follow Futures:
@NatureFutures
go.nature.com/mtoodm

Dave Kavanaugh *is an American science fiction and fantasy author living in the Netherlands. His first novel,* Age of Omicron, *will be released in October. To learn more, visit www.davekavanaugh.com.*

ILLUSTRATION BY OLI WINWARD

Feeling the heat

Earth-observation scientists are forging a way through an uncertain funding landscape to better understand a changing climate.

BY KRISTOPHER JAMES KENT

Steve Montzka, a chemist at the US National Oceanographic and Atmospheric Administration (NOAA) in Boulder, Colorado, is pragmatic when it comes to funding his research. "Throughout the 27 years that's been my career here, it's always been an issue, trying to maintain our capabilities and justifying our existence — and rightly so," he says.

Montzka leads a NOAA project that tracks the atmospheric levels of ozone-depleting substances, and made headlines this May after he and an international research team discovered an uptick of up to 38% since 2012 in global emissions of CFC-11, one of the chemicals banned under international law since 1987 (see S. A. Montzka *Nature* **557**, 413–417; 2018).

After narrowing down the possible causes of these results — eliminating factors such as a natural shift in the atmosphere or the demolition of CFC-containing buildings — the researchers concluded that new CFC-11 is being emitted somewhere in eastern Asia. In July, the Environmental Investigation Agency, a London-based non-profit organization, identified one culprit by reporting widespread use of CFC-11 in the Chinese construction industry, at a level that the organization says could explain the hike in emissions.

Montzka believes that discoveries such as that of his team prove the merits of atmospheric research.

But that isn't enough reason to rest easy. Even before Donald Trump took over the United States' presidency in January 2017, his often-hostile approach to global environmental policy, and corresponding proposed budget cuts to US environmental agencies such as the Environmental Protection Agency (EPA), had alarmed researchers.

The budget that Trump proposed for fiscal year 2018 would have reduced the EPA's budget by more than 30% to US$5.7 billion, before Congress stepped in to preserve the agency's funding. Trump's 2019 budget proposes a similarly huge cut to the EPA's budget.

Montzka explains that declining resources even before Trump's proposed cuts, such as the 2017 rollback from six to four NOAA Atmospheric Baseline Observatories, which carry out solar and atmospheric measurements, make it tougher to analyse the abundant data available. Further cuts would only exacerbate this problem. "News you can use: that's what we'd like to provide," Montzka says. "Substantial budget cuts, if they do occur, would make that more difficult."

What Montzka finds most worrisome, however, is "the waning interest of young scientists. Less funding from government programmes to government science agencies ▶

and universities means many fewer young scientists pursuing the field. Young scientists today are not turning to climate research in the US because of the negative outlook.”

Earth observation is especially vulnerable to financial cuts, owing to the sheer expense of the equipment: a typical weather satellite might cost in the region of $290 million.

But climate scientists are finding ways to navigate the uncertain funding landscape. And they are also finding more opportunities for data-driven experimentation, made possible by new technologies and high-performance computing.

OBSERVING FROM A DISTANCE

Boris Quennehen is an atmospheric scientist at the French start-up Plume Labs in Paris, which is producing a personal, mobile-app-linked pollution monitor for the general public. Despite the company’s focus on the personal, Quennehen believes that government-funded institutes have an indispensable role in environmental observation, one that companies such as his cannot be expected to fill in the wake of dissipating budgets. “The satellite data that agencies such as NASA provide for free are absolutely mandatory if we want to continue to understand the changing climate as our planet evolves,” he says.

Kristie Ebi, a public-health scientist at the University of Washington in Seattle, says that underfunded national meteorological services, especially those from low income countries, sometimes try to make up the deficit by charging researchers exorbitant amounts of money to access their data. She recalls a time when her research group was quoted more than $100,000 for ten years of weather data. For this reason, she says, and in areas lacking other means of data collection, satellites are incredibly important.

Satellite technology has been fuelling Earth-observation studies since the launch of Sputnik 1 in 1957. Now, hundreds of satellites comprising systems such as the European Union’s Copernicus, and NASA’s Earth Observing System, collect and offer data on a host of measurements including sea and land temperature, atmospheric gases, air quality and sea-level changes. These data are typically given to the research community through open-source platforms. Why? To help with understanding, says European Space Agency (ESA) astronaut and geophysicist Alexander Gerst.

Speaking to *Nature* from aboard the International Space Station (ISS), Gerst says that improving understanding of complex Earth systems is the single most important ingredient for the success of future environmental research. “It’s like a machine with a thousand buttons and levers; we do not know what this machine does,” he says, “and how do we fix it, right? Can we just turn another two or three buttons, or is this going to make it better, or worse? We cannot answer this if we do not have any understanding of this climate system — things are going to get worse whatever we do.”

Climate-change research is a focal point for ESA, says Gerst, something that will “of course” continue in future. Earth observation is an ongoing mission aboard the ISS, including studies into developing new technologies for satellites. This year ESA celebrated a significant milestone in its Earth monitoring. On 11 July, the agency announced that Sentinel-5P, its latest satellite and the first Copernicus satellite dedicated to atmospheric measurement, has produced its first data: air-pollution maps showing a range of trace gases that affect air quality. Using a cutting-edge spectrometer, Sentinel-5P can detect multiple air pollutants at city-level resolution.

PRIVATE AIR PURITY

Quennehen and his colleagues at Plume Labs are attempting to use their palm-sized pollution-detection device, dubbed Flow, which is nearing the end of its development and production phase, to educate the public. “People really do want to know what they breathe, but today they can’t access that data by themselves,” he says. “Ultra-local data is missing today in the existing monitoring networks.”

Plume Labs is collaborating with research groups across Europe, in particular the atmospheric research laboratory CNRS-LISA in Paris, which provides the company with lab space and facilities. Quennehen says that, in return, Plume could help the CNRS, France’s national scientific research organization, to save costs in mapping areas where data collection is challenging. “It’s not easy to find funding for such projects,” he explains. “Flow could help them to map pollution in remote areas for quite low prices.”

PEOPLE REALLY DO WANT TO KNOW WHAT THEY **BREATHE**, BUT TODAY THEY CAN’T ACCESS THAT DATA BY **THEMSELVES.**

Because of the portability of the device, Plume is also courting interest from epidemiologists, who are keen to understand the pollution exposure of their study participants.

Another collaboration between a private company and researchers is taking place in London. The company Air Monitors, based in Tewksbury, UK, alongside researchers at the University of Cambridge’s chemistry department, the National Physical Laboratory in Teddington and other partners, is installing 100 air-quality monitors across London. The monitors will connect wirelessly with a database, and will operate alongside Google Street View cars to draw a map of pollution on the city’s streets in close to real time.

Other private ventures are also taking advantage of greater public awareness to market new products to remove, measure and contextualize pollutants in the air. Molekule, in San Francisco, California, for example, says that its air purifier destroys indoor pollutants through a chemical process of photoelectrochemical oxidation. Foobot, based in San Francisco and Luxembourg, offers a home detector that tracks air pollutants and visualizes their concentration by the intensity of its LED display. And the Shit, I Smoke! app draws an equivalence between air quality in a city and cigarette consumption. (A day in Beijing? Four cigarettes.)

THINKING OUTSIDE THE BOX

Sean Beevers, an air-quality researcher at King’s College London, and his team have produced pollution maps at an even higher resolution than those from Sentinel-5P, using models that combine road-vehicle data with an understanding of atmospheric processes.

“Satellite measurements are useful in some respects, but have a high level of uncertainty” when it comes to making accurate measurements, says Andrew Beddows, a member of the team. In addition, the Sentinel-5P can measure pollution only during the daytime and with clear skies, he points out. Nevertheless, Beevers is hoping to apply his team’s modelling prowess to parts of the world that lack sufficient traffic data, by using satellite imagery to count

MANJUNATH KIRAN/AFP/GETTY IMAGES

Houses destroyed by a landslide in Kerala, India, after exceptionally heavy rain in August.

vehicles and estimate their speed.

Like Quennehen and his colleagues, Beevers says that detailed air-quality maps will be useful in estimating an individual's exposure to pollution, and can power investigations into air quality and health. But he says that he retains a healthy degree of scepticism regarding technology such as Plume's. "Whether the wearable tech things will add to that, I don't know," he says. "So far, it's not provided everything it's promised to, but things are changing quite rapidly in that area, so you never know."

HIGH-PERFORMANCE SCIENCE

A key driver in understanding how our planet works is the development of high-performance computing, says Ajaya Ravindran, a climate scientist at New York University Abu Dhabi.

A climate researcher for almost 20 years, Ravindran, who hails from India, says that over the course of his career "improvement in computational power is the most important change I can see". Rainfall in particular, Ravindran says, is enormously taxing to model accurately.

Early this year, he and his colleagues predicted a 45% decline in the atmospheric low-pressure systems that drive monsoon rainfall, accompanied by a directional shift that threatens to increase the risk of extreme weather events in India (see S. Sandeep *et al. Proc. Natl Acad. Sci. USA* **115,** 2681–2686; 2018). One occurred this summer, when floods killed nearly 400 people and displaced more than 1 million after heavy rain in Kerala.

Ravindran and a colleague have also predicted that the monsoon season would shorten by about 11 days (C. T. Sabeerali and R. S. Ajayamohan *Clim. Dyn.* **50**, 1609–1624; 2018). "A changing climate is crucial — it's crucial everywhere in the world, but in India? Even more so," he says.

Ravindran believes that short-termism is at least partly to blame for the relaxed attitudes of policymakers when it comes to climate change, but he says that the global mood is shifting. "Trump is an anomaly," he says, adding that the governments of Canada and some EU countries are increasingly prioritizing the issue of climate change. But, he says, there is still work to be done.

Ebi says that linking climate change to public health is key to combating short-termism. In May, she and her colleagues reported that increased carbon dioxide levels in this century will reduce the nutritional quality of rice, threatening the health of 600 million people in low-income southeast Asian countries (see C. Zhu *Sci. Adv.* **4**, eaaq1012; 2018).

"Not just America, but every country needs to put more money into climate change," says Ravindran. "But the US is especially sad to see. Some of the major observation networks are there. They're the ones maintaining the best satellites, and some of the best scientists are also there."

Montzka agrees: "Climate research in the US runs a very real risk in the future of falling from being the best in the world, to soon becoming irrelevant."

As the public start to feel the heat, their responses will be reflected in elections, says Ravindran. "People have started feeling it. Even in Kerala, one of the highest-rainfall provinces in India, there is water scarcity nowadays. Every year it's becoming hotter, and more real." ■

Kristopher James Kent *is a freelance journalist in Liverpool, UK.*

Keeping an eye on climate change

Orbiting hundreds of kilometres above the Earth's surface, satellites launched by the Japan Aerospace Exploration Agency (JAXA) are at THE FOREFRONT OF CLIMATE RESEARCH

The Greenhouse gases Observing SATellite-2 (GOSAT-2) in orbit

© JAXA

Climate change caused by humans threatens to have serious consequences for populations around the world through its impact on both terrestrial and marine ecosystems. More frequent and extreme weather events, including heat waves, droughts and flooding, will exact significant tolls for generations to come.

Understanding how climate is influenced by greenhouse gases, such as the main drivers of global warming, carbon dioxide and methane, is essential to tackle the problem. But this presents an enormous scientific challenge, since it requires monitoring the factors that influence climate at local, regional and global scales over decades. That is where satellite observations play a vital role.

Using state-of-the-art satellite technology, the Japan Aerospace Exploration Agency (JAXA) is at the vanguard of climate science and research. “Continuous and detailed monitoring by satellites allows us to observe changes at the global scale,” says JAXA's president, Hiroshi Yamakawa. “They provide scientists with the data to better understand the changes occurring in the Earth's climate.”

SATELLITES JOIN A GLOBAL NETWORK

An extensive network of atmospheric, oceanic and land-based monitoring systems — including weather stations, buoys, ships, balloons and aircraft — collect data on more than 50 parameters that influence the planet's climate, including sea ice coverage, forest cover and precipitation. JAXA, the National Institute for Environmental Studies (NIES) and the Japanese Ministry of the Environment (MOE) jointly launched the first satellite dedicated to monitoring the sources and sinks of greenhouse gases in 2009. Since then, greenhouse-gas observation satellites have joined this global network, becoming an integral tool in climate research.

The Greenhouse gases Observing SATellite (GOSAT) — a collaboration between JAXA, NIES and MOE — has provided invaluable data for verifying the reductions in greenhouse-gas emissions agreed under the Kyoto Protocol to the United Nations Framework Convention on Climate Change, adopted in 1997.

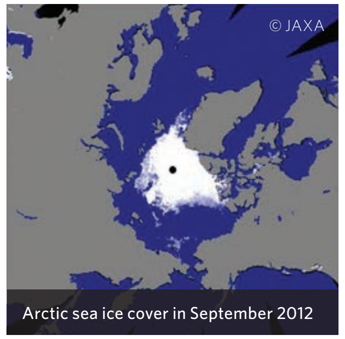
“We're now integrating data from GOSAT with observations from aircraft, ships and land-based monitoring systems,” explains Nobuko Saigusa, director of the Center for Global Environmental Research at NIES. “This is leading to significant improvements in monitoring greenhouse gases emitted by man-made and natural systems.”

TRACKING WATER AND ICE FROM SPACE

Of the many factors that influence climate, the Earth's water cycle is particularly important. To understand how climate change impacts the global water cycle, JAXA is driving the Global Precipitation Measurement (GPM) mission and the Global Change Observation Mission (GCOM), which consists of two satellites, GCOM-W and GCOM-C.

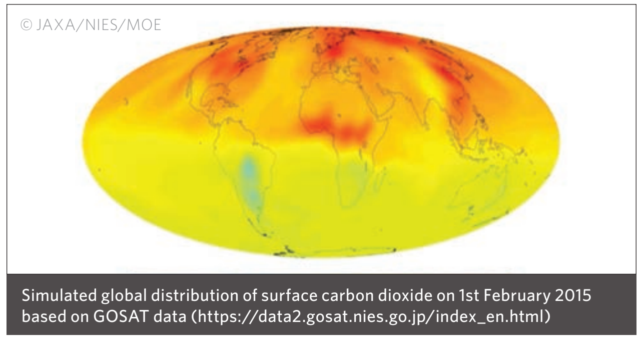
JAXA's president, Hiroshi Yamakawa

Nobuko Saigusa, director of the Center for Global Environmental Research at the National Institute for Environmental Studies (NIES)



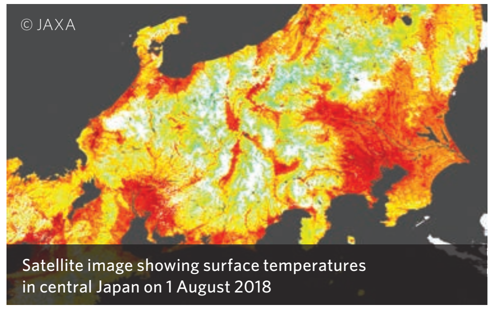
© JAXA

Arctic sea ice cover in September 2012



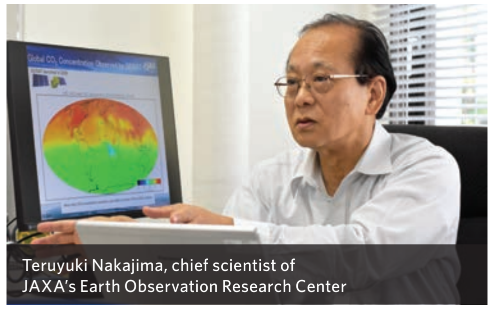
© JAXA/NIES/MOE

Simulated global distribution of surface carbon dioxide on 1st February 2015 based on GOSAT data (https://data2.gosat.nies.go.jp/index_en.html)



© JAXA

Satellite image showing surface temperatures in central Japan on 1 August 2018



Teruyuki Nakajima, chief scientist of JAXA's Earth Observation Research Center

Since its launch in 2012, GCOM-W has been monitoring precipitation, sea ice, sea surface temperature, and soil moisture. Using the world's largest scanning antenna — the Advanced Microwave Scanning Radiometer 2 (AMSR2) — GCOM-W is able to cover over 99% of the Earth's surface in just two days. In September 2012, AMSR2 measured the smallest extent of Arctic sea ice on record. And in July 2017, AMSR2 captured dramatic images of a giant iceberg splitting from the Larsen C ice shelf in Antarctica. At nearly 6,000 square kilometres — more than twice the size of Luxembourg — and weighing over a trillion metric tonnes, the calving event reduced the ice shelf by 10%. Launched by JAXA in 2017, GCOM-C is monitoring aerosols, clouds, vegetation and temperatures over land and oceans, which is helping to enhance the accuracy of forecasting environmental changes.

GPM is a joint mission between JAXA and NASA to measure global rain and snow with high frequency and high accuracy by using multiple satellites. Its core observatory, which carries a dual-frequency precipitation radar and GPM microwave imager, was launched in 2014. By mapping global precipitation from the Arctic to Antarctic, GPM is providing greater insights into the Earth's weather and climate change studies.

“Modern satellites collect huge amounts of data from across the oceans, the Arctic and other sparsely populated zones, which are difficult for humans to monitor,” notes Teruyuki Nakajima, chief scientist of JAXA's Earth Observation Research Center.

THE FUTURE OF EARTH OBSERVATION

This year, JAXA, NIES and MOE will jointly launch GOSAT-2, the successor to GOSAT. Carrying advanced pointing capabilities that allow it to avoid cloud cover, GOSAT-2 will measure carbon dioxide, methane and carbon monoxide levels with greater accuracy across a broader range of locations, including industrial and densely populated areas. Its data will contribute to the first global greenhouse-gas inventory required under the Paris Agreement, which is scheduled for release in 2023.

Scheduled for launch, the EarthCARE is an international collaboration between JAXA, Japan's National Institute of Information and Communications Technology (NICT) and the European Space Agency (ESA). Japan is contributing one of its on-board instruments, the cloud profiling radar. It will monitor the structures of clouds and aerosols in three dimensions and promises to significantly reduce uncertainties in estimates of the Earth's radiation budget for better climate change prediction.

JAXA is also leading international collaboration on the deployment of Earth-observation technologies, through agreements with NASA and ESA, as well as the French and German space agencies. “Reliable data that can be analysed and transformed by scientists into useful information for policymakers cannot be achieved by one agency,” says Yamakawa. “It requires collaboration between space agencies around the world.” ■

JAXA
Explore to Realize
http://global.jaxa.jp/